Plasmid genomic studies of the large phytosphere associated plasmid pQBR103

Adrian James Tett
Student number: 971702026

A thesis submitted for the degree of Doctor of Philosophy

Cardiff University

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ......A. Tetty.. (candidate) Date..20/06/07.....

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD

Signed......A. Tetty.. (candidate) Date....20/06/07....

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed ......A. Tetty.. (candidate) Date...20/06/07...

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ...A. Tetty... (candidate) Date....20/06/07..

Thesis Abstract

pQBR103 is a representative of the large pQBR collection of ecologically well characterised self-transmissible and mercury resistant plasmids. These plasmids constitute five genetically distinct groups (I-V) and were isolated from the phytosphere of a number of plant species at a single geographical location. Of the collection pQBR103 (group I), which is known to confer a seasonal fitness benefit on a pseudomonad host, is the best studied.

Determination and annotation of the pQBR103 sequence, in this thesis, has revealed it is large at 425 kb, and contains no 16S rRNA or other host "housekeeping" genes. It also revealed the likely mechanisms of plasmid replication and stability, but not of plasmid transfer. While traits of possible ecological function were identified, 80% of coding sequences could not be ascribed a function and 59% of the total coding capacity of pQBR103 constituted novel orphan genes. Therefore, the ecological function of this plasmid remains largely enigmatic. Having created an online database, of all publicly available complete plasmid genomes, the lack of functional assignment to pQBR103 may reflect the biases in plasmids previously chosen for sequencing.

Building on the legacy of the plasmid sequence, genomic investigations were undertaken. *In vitro*, proteomic studies demonstrated pQBR103 is tightly regulated and affects host chromosomal responses to the environment. Diversity within the pQBR collection (groups I, III and IV) was assessed by PCR surveys and a custom pQBR103 microarray. Intergroup comparisons revealed no shared homology between group I and groups III or IV, other than common *mer* genes, suggesting independent evolutionary histories. Intragroup comparisons identified genetic instability and revealed a common genetic core, which contains the genes putatively responsible for stability and replication. These features are not clustered and there are relatively few repeat sequences, both of which contradict the modular paradigm described for other plasmids.

Acknowledgements

For all those who helped make this thesis possible, whether that help was scientific advice or just a cold pint after a hard days work, I offer my eternal gratitude.

I acknowledge NERC for funding this PhD and CEH Oxford for providing me a home in which to study. In particular I would like to thank all my supervisors Mark for his wealth of knowledge, Dawn for advice and all things Bioinformatics (I wouldn't have got here without you) and John for his invaluable advice all things Cardiff related. I would also like to give a special thanks to Sarah for continuous support and supervision, I owe a lot of this thesis to her.

Credit must also be given to all the members, past and present, of the Molecular Microbial Ecology group, Molecular Evolution and Bioinformatics group and the NERC Environmental Bioinformatics Centre (NEBC), Oxford crew. In particular I thank Andy for advice and field experiments also, Bruce, Lena and Kate for help with processing the samples. I am sorry the experiment never took off.

Other thanks go to the Sanger Centre, Cambridge, for making pQBR103 possible. Lori and Nigel from the Dunn School of Pathology, Oxford University, for the microarrays and letting me invade their laboratory and Bela from NEBC for all her help in interpretation. I also thank Steph and Xiao Hong for help with the proteomic experiments.

Thanks must also be given to family and friends for their support, I am not going to list you - you know who you are. OK, Phil can have mention he has heard me moan enough so he at least deserves that.

My greatest thanks of all goes to Kate, for her support and for putting up with more than anyone should have to. Without her this thesis would not be a reality without acknowledging that I would be without her.

Table of Contents

List of abbreviations

bp	Base pair
BSA	Bovine Serum Albumin
°C	Degrees centigrade
CDS	Coding Sequence
cfu	Colony forming unit
Chaps	3-[3-Cholamidopropyl)-Dimethylammonio]-1-Propane Sulfonate
CGH	Comparative genomic hybridisations
COG	Clusters of Orthologous groups
CTAB	Hexadecyltrimethylammonium bromide
DDBJ	DNA Data Bank of Japan
2-DE	Two dimensional electrophoresis
DTT	Dithiothreitol
DNA	Deoxyribonucleic acid
dNTP	Deoxy Nucleoside Tri-Phosphate
EDTA	Ethylenediaminetetraacetic acid
EMBL	European Molecular Biology Laboratory
Et-OH	Ethanol
g	Grams
g	Gravity
GOLD	Genomes OnLine Database
HGP	Horizontal gene pool
HGT	Horizontal gene transfer
IVET	In vivo expression technology
J	Joules
Kb	Kilobase
KDa	Kilodaltons
LB	Luria Bertani broth
LBA	Luria Bertani Agar
LMP	Low melting point
µg	Micrograms
µl	Microlitre
M	Molar
Mbp	Megabase pair
MIGS	Minimum Information about a Genome Sequence
mg	Milligrams
MGE	Mobile genetic element
mL	Millilitres
mM	Millimolar
mRNA	Messenger ribonucleic acid
m/s	Metres per second
MS	Mass spectrometry
NaAc	Sodium acetate
NCBI	National Centre for Biotecnology Information
NDSB-221	3-(1-Methylpiperidinium)-1-propane Sulfonate
ng	nanogram
orf	Open reading frame
PBS	Phosphate Buffered Saline

PCA	Plate Count Agar
PCR	Polymerase Chain Reaction
pg	Picogram
PGD	Plasmid Genome Database
pM	Picomolar
PMA	Phenylmercuric acetate
PMF	Protein Mass Fingerprint
PSA	Pseudomonad Selective Agar
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic acid
s	Seconds
SDS	Sodium Dodecyl Sulphate
SSC	Saline Sodium Citrate (Standard Sodium Citrate)
TBP	Tributylphosphine
TE	Tris-HCL EDTA
UV	Ultra violet
V	Volt
V-hr	Volt hour
vol	Volume
WA	Water agar
wt	Weight

Chapter 1: Thesis Introduction and Aims

Section 1.1 Horizontal gene transfer (HGT) and mobile genetic elements (MGE)

The role bacteria and other micro-organisms play in shaping and ultimately sustaining life on earth is becoming increasingly apparent. Indeed, many fundamental processes of life are attributable to these organisms (Smets and Barkay, 2005). In a complex environment such as soil, bacteria and other microbes are responsible for processes such as nutrient cycling which is essential to plant survival, the primary producers of energy on earth. Early models of bacterial adaptation, evolution and speciation were based on clonality and periodic selection (Gogarten et al., 2002). It is now accepted that such models based on genetic drift alone are insufficient to explain the large metabolic diversity displayed by bacteria (Ochman et al., 2000). A better explanation for genetic innovation and functional genomic plasticity is horizontal gene transfer (HGT).

Clonality arises from genes passing mother to daughter (vertical transmission). HGT is the intra and interspecific exchange of genes within generations. This goes someway to compensating for the clonal lifestyle of bacteria (Smets and Barkay, 2005) i.e. it purges the deleterious mutations that are predicted by Muller to accumulate in a clonal population of finite size, Muller's ratchet (Muller, 1932). Genes that transfer horizontally are frequently contained on mobile genetic elements (MGEs). MGEs can be defined as "segments of DNA that encode enzymes and other proteins that mediate the movement of DNA within genomes or between bacterial cells" (Frost et al., 2005). Such elements include plasmids, bacteriophage, transposons, integrons, genomic islands and frequently combinations of these elements. The genes they encode and disseminate represent the horizontal gene pool (HGP).

1.1.1 Mechanisms of HGT

There are essentially three predominant mechanisms for HGT into a new host. These are transformation (Griffith, 1928), transduction (Lederberg et al., 1951) and conjugation (Lederberg and Tatum, 1946). Of the three methods only transformation is not mediated by MGEs but relies on a phenomenon known as natural competence, in which bacterial cells actively take up DNA. Transduction is a mechanism of transfer mediated by bacteriophages which package and transfer DNA as part of a lifecycle, and sometimes accidentally package and transfer non-phage DNA between cells. Unlike both transformation and transduction, conjugation involves cell to cell interaction via a protein pilus before the DNA is transferred to the recipient. This process is complex involving multiple genetic constituents that are often co-localised for co-expression (see section 1.2.1.3.1) on conjugative plasmids, transposons and integrated chromosomal elements. A schematic summarising HGT transfer mechanisms is given in Figure 1.1. Plasmid transfer is further introduced in section 1.2.1.

Transfer of DNA alone is not sufficient for a successful HGT event. The newly acquired DNA must be stabilised and maintained in its new host by replicating thus enabling it to be inherited vertically. This is either achieved by integrating into the host chromosome (e.g. the *Escherichia coli* bacteriophage lambda) or by forming an autonomously replicating extrachromosomal element (e.g. plasmids but also some non integrative phage e.g. P1) (Lobocka et al., 2004). There are two types of recombination; homologous and non-homologous recombination, also referred to as legitimate and illegitimate recombination (Rocha, 2003). Homologous recombination requires extensive homology between the newly acquired DNA and the recipient's chromosome for pairing and strand exchange. This can occur independently from MGEs i.e. as is the case with transformation. This recombination is sometimes referred to as general homologous recombination to distinguish it from site specific homologous recombination, which occurs between short specific homologous sequences that some MGE contain to enable their integration into other DNA elements (Lengeler et al., 1999). Both homologous and site specific recombination are largely dependent on host encoded recombination systems (Rec). On the other hand illegitimate recombination is Rec independent and requires only very short regions of homology (5-10 bp).

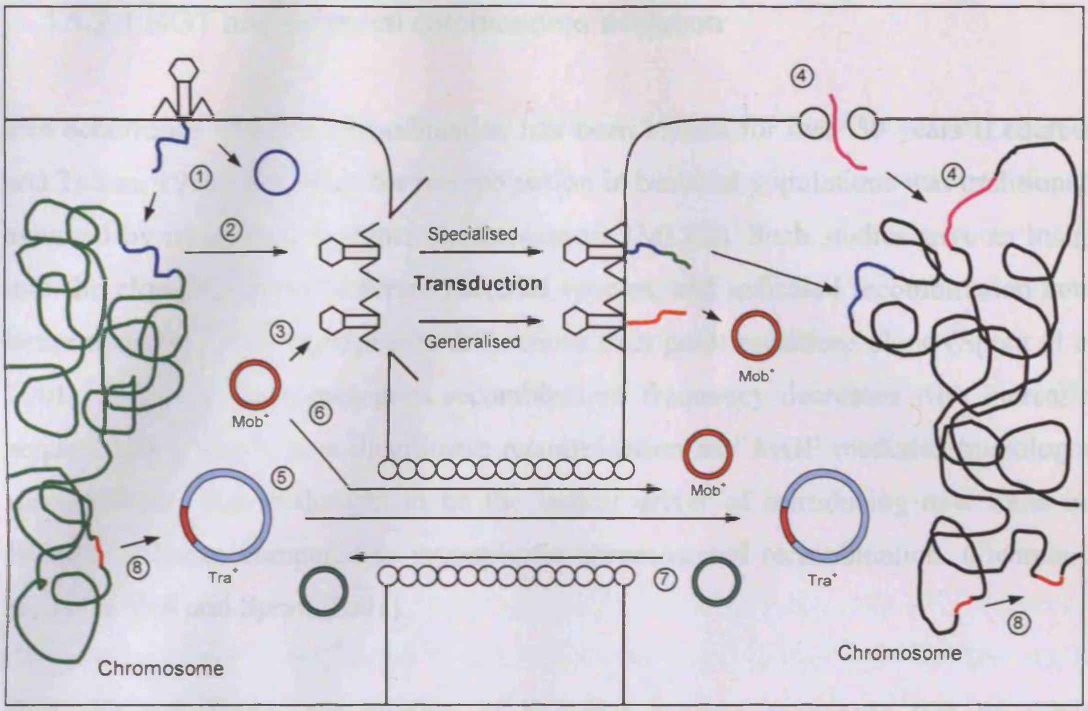


Figure 1.1. Summary of intra and interspercific gene transfer. Figure adapted from Frost et al., 2005
1) Bacteriophage infection, inserted DNA (blue) becomes incorporated into the chromosome as a prophage, or forms an independent replicating element 2) Specialised transduction, incorrect prophage excision causes prophage (blue) as well as Chromosomal DNA to become packaged. 3) Generalised transduction, non-phage DNA becomes packaged in phage capsids, which in this example is plasmid DNA (orange). 4) Transformation, DNA from the extracellular environment (pink) is taken up into the cell and in this example incorporated into the host chromosome. 5) Self mediated plasmid transfer by conjugation. 6) Mobilisable plasmid transfer mediated by conjugative apparatus of helper strain. 7) Retrotransfer or retromobilisation mediated by conjugative apparatus of helper strain. 8) Transposon mediated transposition.

1.1.2 Consequences of HGT

HGT was first recognized as far back as the 1920's when virulence determinants were first transferred between pneumococi in mice by transformation (Griffith, 1928). However, it was with the emergence of antibiotic resistant pathogens on a world wide scale in the 1950s and 60s that truly highlighted the importance of HGT via mobile genetic elements (Summers, 1996). It soon became apparent that MGEs are near ubiquitous throughout the natural environment and bacterial systems (Bergstrom et al., 2000). Also that these elements encode many proteins responsible for the varied traits including xenobiotic degradation, heavy metal resistance, nitrogen fixation, the synthesis of plant growth regulators, toxin production, and animal pathogenicity determinants (Summers, 1996; Bailey et al., 2001). Nevertheless, it has only been relatively recently that the true extent of HGT in the short term adaptation and the longer term evolution has been appreciated.

1.1.2.1 HGT and bacterial chromosome evolution

The occurrence of DNA recombination has been known for over 50 years (Lederberg and Tatum, 1946). Evidence for recombination in bacterial populations was traditionally assessed by multi locus enzyme electrophoresis (MLEE). Such studies gave an insight into the clonality within different bacterial species, and indicated recombination could better account for lineage-specific differences than point mutations alone (Spratt et al., 2001). However, as homologous recombination frequency decreases with increasing sequence divergence, it is illegitimate recombination and MGE mediated homologous recombination that is thought to be the largest driver of introducing new traits into bacterial lineages compared to intraspecific chromosomal recombination. (Ochman et al., 2000; Feil and Spratt, 2001).

Over the last decade the number of bacterial genome sequences that have been deposited in the public databases has increased exponentially. Examination of the sequences has given better estimates of the extent that HGT has played in bacterial evolution. Firstly, by identifying the remnants of MGE which are indicators of HGT. Secondly, as different bacteria vary in base compositions i.e. G+C content, codon and amino acid usage, by the identification of regions atypical to the chromosomal and, therefore, assumed to be horizontally transferred (Lawrence and Ochman, 1997). However, such studies only offer a "snap shot" in time, identifying successful HGT events that have become fixed in the population or represent regions that are yet to be lost. These methods are also likely to underestimate the level of HGT as not all transferred DNA that recombines into the chromosome will display different base compositions. Moreover, transferred DNA that is resident for a sufficient period of time will become ameliorated, taking on the compositional characteristics of its host (Lawrence and Ochman, 1997). Regardless of these caveats, estimates of the proportion of genes acquired horizontally are revealing. *E. coli,* originally thought to be clonal via MLEE studies (Smith et al., 1993) has been predicted to have gained 17.6% of its coding sequences via HGT (Lawrence and Ochman, 1998). Such studies show that although the frequency of recombination may be low, the fact it occurs at all is an important issue in the long term evolution of bacterial species (Levin and Bergstrom, 2000).

1.1.2.2 HGT and bacterial adaptation

In terms of bacterial evolution the role of HGT is becoming clear and is widely accepted. However, the slow process of evolution cannot explain the observed rapid adaptation to perturbations in the environment by bacterial communities in response to environmental stresses e.g. in response to stresses like the increased use of antibiotics or xenobiotics; nor could such a system explain the rapid exploitation of temporal niches, such as seasonal relationships with higher organisms such as plants (Top and Springael, 2003). These observations can be better explained by horizontally transferred DNA that is only transiently associated with the host i.e. not destined to participate in bacterial evolution.

This "transiently" associated DNA that represents the highly dynamic HGP constitutes only a small proportion of a cell's total coding capacity. Nevertheless, its effect on the behaviour of the cell can be wide reaching. Indeed it is the tumour inducing and transfer systems that have made *Agrobacterium tumefaciens* of interest to the biotechnologist, all of which are partitioned in the horizontal gene pool (Wood et al., 2001; Goodner et al., 2001). Similarly, the difference in the lifestyle of *Bacillus anthracis*, *Bacillus. cereus* and *Bacillus. thuringiensis* from pathogen to being of agrochemical important is largely due to extrachromosomal plasmids (Helgason et al., 2000). The importance of plasmids in Bacillus pathogenesis was demonstrated when the transfer of a plasmid that was near identical to that of *B. anthracis* pX01 to *B. cereus* caused a disease resembling inhalation anthrax (Hoffmaster et al., 2004).

1.1.3 Focus of the thesis

Whether HGT goes to "completion" i.e. the genes become fixed in the chromosome if selection is significantly high (Berg and Kurland, 2002), is of interest to the evolutionary biologist. However, it is the more transient nature of horizontally transferred MGEs that is the study of the microbial ecologist (Smets and Barkay, 2005) and is ultimately the focus of this thesis. Additionally, it is suggested that MGEs are to varying degrees interrelated, consisting of combinations of modules that interchange between distant types of MGE (Toussaint and Merlin, 2002). Accepting that at the extremes of MGE classification there is a degree of "blurring", on the most part demarcation is a valid and useful exercise. In this thesis the focus is firmly towards the study of bacterial plasmids.

Section 1.2 Bacterial plasmids

Bacterial plasmids are extrachromosomal elements that replicate autonomously from the host chromosome and are usually circular double stranded DNA molecules, although some are linear. Plasmids vary considerably in terms of size (~1 kb - over 1 Mb) and the functions they encode. Broadly speaking plasmid encoded functions can be divided into two categories, those that may be of potential benefit to host i.e. UV resistance and synthesis of plant growth regulators; and those responsible for plasmid maintenance i.e. replication, partitioning, dimer resolution and conjugative transfer (Bailey et al., 2001; Thomas, 2000). Some of these functions are discussed in chapter 2. In this chapter section, specific mechanisms for plasmid transfer, persistence and diversity in natural environments are addressed.

1.2.1 Bacterial plasmid transfer mechanisms

Earlier, the predominant mechanisms for HGT were introduced (1.1.1). Plasmids are amenable to transfer by all three of the methods listed; transformation, transduction and conjugation. Although, all methods are relevant they are not all of equal importance.

1.2.1.1 Transformation

The process of transformation and its relevance to bacteria is dependent on a number of factors. These include the environmental conditions the bacterium is experiencing and the host's degree of natural competency. Not all bacteria are competent yet it has been demonstrated for many different bacteria with many different physiologies (Lorenz and Wackernagel, 1994). In some environments free DNA has been shown to be relatively stable and relatively common (Davison, 1999). For example even in blood serum free plasmid DNA has been shown to persist for a long enough period to be available for uptake by infectious bacteria (Rozenberg-Arska et al., 1984). Additionally, Williams et al. (1996) demonstrated plasmid acquisition *in situ* via transformation in aquatic environments.

1.2.1.2 Transduction

Transduction can be divided into two types, specialised and generalised both of which involve the packaging of non-phage DNA. Specialized transduction is an error of the lysogenic life cycle of the phage. When the prophage is induced excision is carried out incorrectly and the genes surrounding site of phage integration become packaged and transferred along with the phage DNA (Lengeler et al., 1999). As phage incorporation into plasmids is low, and if occurs is unlikely to transfer a full complement of plasmid genes, the consequences of specialised transduction in plasmid transfer is most likely to be minimal. Unlike specialised transduction, no phage DNA is necessarily transferred in generalized transduction. Here an error in the assembly stage, when phage capsids are being packaged, allows for the non-discriminatory packaging of DNA of the correct size into the capsid. Therefore, it is generalised transduction that is most likely to be of consequence in plasmid transfer. However, as bacteriophages are generally of narrow host range it is unlikely that this method plays as prominent a role in gene transfer compared to transformation and conjugation. Regardless, phage concentrations can be large in freshwater, marine and terrestrial environments (Edwards and Rohwer, 2005) and transfer of plasmids via phage in the natural environment has been demonstrated (Jiang and Paul, 1998), suggesting their role is not insignificant.

1.2.1.3 Conjugation

Two types of conjugative transfer are available for plasmids; self conjugation whereby the process is encoded on the plasmid itself, and mobilisation whereby the plasmid is parasitic, relying on the transfer mechanisms encoded on other plasmids. Regardless of whether mediated by self or not, conjugation is the dominant mechanism for plasmid spread in the environment, exemplified by early exogenous plasmid isolations (for a description of environmental plasmid isolation see box 1) (Bale et al., 1987). Example studies include: Bale et al., (1988); Lilley et al., (1994); Top et al., (1994); Dahlberg et al., (1997), for review see Davison (1999). Due to the importance of conjugation in driving HGT the mechanism for this is considered further.

Box 1 Environmental plasmid isolation techniques

Endogenous isolation. The traditional method of plasmid isolation is the endogenous method whereby plasmids are isolated in their natural culturable hosts. Bacteria are first grown in pure culture possibly for a selectable phenotype (e.g. heavy metal resistance) than subsequently screened for the presence of plasmids by DNA extraction (Smalla et al., 2000).

Exogenous isolation. In exogenous isolation plasmids are not isolated in their natural host but captured into recipient bacteria directly from the environmental sample. There are two types of exogenous isolation, **biparental** (Bale et al., 1987; 1998) and **triparental** (Hill et al., 1992). In biparental matings plasmids are isolated by their ability to transfer (either independently or mobilised by a helper plasmid) from the indigenous bacterial (donor) community (environmental sample) into a marked recipient. Transferred plasmids must contain a selectable marker (e.g. mercury resistance). The original donor population and plasmid free recipients are counter selected by using a combination of the plasmid encoded selectable marker and recipient chromosomal marker (e.g. rifampicin).

As the name suggests in triparental mating there are three elements, two donors and a recipient. Donor 1 is the indigenous bacterial community and donor 2 is a known strain containing a plasmid encoding a selectable marker, this plasmid is mobilisable but not self transferable. The recipient, as in biparental mating, is a marked strain. Plasmids are isolated based on their ability to self transfer from the donor 1 population into the strain containing the mobilisable plasmid (Donor 2). Transfer of the environmental plasmid to the recipient is accompanied by mobilising the marked mobilisable plasmid. Donors and plasmid free recipients are counterselected using the selectable markers encoded by the recipient and mobilisable plasmid. The major difference between the two exogenous isolation methods is the selection. In biparental mating the environmental plasmid contains the selectable marker, in triparental matings plasmids are captured by their ability to mobilise the mobilisable plasmid. Both exogenous methods allow the isolation of environmental plasmids irrespective of the culturability of the original host.

1.2.1.3.1 The mechanism of conjugative transfer

Bacterial conjugation was first demonstrated by Lederberg and Tatum (1946), when prototrophic growth was obtained after the incubation of mixed auxotrophic *E. coli* mutants. Although at the time a full explanation of the result could not be given, subsequent studies demonstrated that the transfer was dependent on cell to cell contact and was under the direction of genes encoded on a plasmid; the F plasmid. It is the continued study of this plasmid over the years which have made this one of the best understood conjugative transfer systems. Other Gram negative bacteria with well characterised systems include the broad range IncPα and β Birmingham plasmids (Pansegrau et al., 1994; Thorstead et al., 1998) (for a description of plasmid Inc groups see box 2, page 13) and the *A. tumafaciens* pTiC58 plasmids (Wood et al., 2001; Goodner et al., 2001). Plasmid conjugative systems are not only restricted to plasmids of the Gram negative bacteria, they have also been identified in Gram positive bacteria. Examples of such plasmids are pADP1 of the enterococci and plasmid pIJ101 of streptomycetes (Lengeler et al., 1999).

Box 2 Incompatibility testing of plasmids

Incompatibility for decades has been employed as a formal way to classify plasmids (Smalla et al., 2000). Plasmid incompatibility is the inability of two plasmids to coexist within a single cell line (Summers, 1996). This inability arises primarily due to interference of replication and control. This interference is due to the similarity of these functions between the two plasmids which is correlated at the DNA level (Smalla et al., 2000) i.e. plasmids with a common ancestry will have similar replication regions and thus will be incompatible and assigned to the same incompatibility (Inc) group.

Traditionally plasmids were typed by following the fate of the unknown plasmid against plasmids of known replication type in a cell line. However, the results of from such studies are difficult to interpret e.g. due to some plasmid containing more than a single minimal replicon and because the observed incompatibility may arise because of similar plasmid partitioning functions not replication functions. More recently DNA probes for known incompatibility groups developed by Couturier (Couturier et al., 1988) have been used for typing the incompatibility groups of unknown plasmids by hybridisation techniques. However, the best way of assessing the evolution relationship between plasmids is to combine *in vitro* methods with *in silico* analysis, comparing the nucleotide sequences of the plasmids directly (where the nucleotide sequence of the plasmid/minimal replicon has been determined).

Comparisons of characteristics of the conjugative transfer systems in the Gram negative bacteria have indicated that they are diverse both in terms of genes they encode and their operon structures (Thomas, 2000). Nevertheless, key components of these systems are common to all, suggesting there are common mechanistic principles to these systems (Christie and Vogel, 2000; Schroder and Lanka, 2005). These common principles will be introduced further, particularly with respect to Gram negative systems.

Plasmid conjugation can effectively be divided into two processes. Firstly, mating pair formation (Mpf) that establishes cell-to-cell contact of the donor and recipient. Secondly, DNA transfer and replication functions (Dtr) responsible for the processing of DNA during conjugation (Schroder and Lanka, 2005). These two processes are generally interlinked by a coupling protein (CP). Although these processes are distinct the genes encoding them are not necessarily so. These have been shown to be clustered in a single operon (F plasmid) as well as in distinct clusters (IncPα) (Lengeler et al., 1999). Despite the difference in operon structures a common feature of both the IncP and F plasmids, as well as other characterised systems including that of R388 and pTi, is that the genes of the Mpf system show homology and group with Type IV secretion systems identified in bacterial chromosomes. Such systems are functionally diverse in bacteria, being involved in DNA release and uptake as well as in protein translocation in some pathogenic bacteria. Examples of some Type IV systems are given in Table 1.1.

Function	Bacterium/plasmid name	System name	Reference
DNA transfer	F plasmid (IncF1)	Tra	Lawley et al., 2003a
	RP4 (IncPα)	Trb	Pansegrau et al., 1994
	R751 (IncPβ)	Trb	Thorsted et al., 1998
	pKM101 (IncN)	Tra	Pohlman et al., 1994
Effector protein translocation	*Helicobacter pylori*	Cag	Bourzac and Guillemin, 2005
	Bordetella pertussis	Ptl	Weiss et al., 1993
DNA uptake	*H. pylori*	ComB	Karnholz et al., 2006

Table 1. Type IV secretion systems. Examples of functionally diverse Type IV secretion systems that share homology at the sequence level. Comprehensive or recent review of the system is given as reference.

The first step in the model of bacterial conjugation and Mpf is attachment to a potential recipient and determining whether it is able to mate. This attachment is established by sex pili, the morphology of which can vary between bacteria. The pili encoded by the F plasmid are long and flexible compared to that of IncP pili which are shorter and rigid (Bradley et al., 1980; Lawley et al., 2003b). The environment in which the plasmid transfers could explain these morphological differences. The F plasmid transfers more efficiently in liquid media compared to the IncPα plasmids such as RP4 that transfers more efficiently on solid supports. There is a hypothesis that incP plasmids originate in soil bacteria. Therefore, pilus adaptation to optimise transfer on solid surfaces is a reasonable explanation. In contrast, the F plasmid, found free in liquid environments as well as attached to solid surfaces, is adapted to transfer in both (Bradley et al., 1980; Lawley et al., 2003b).

Attachment of the sex pilus to the potential recipient is not fully understood but is generally dependent on two factors. First the recognition of specific structures on the recipient's outer membrane that could be a membrane protein or lipopolysaccharides; Secondly, the absence of elements that prevent transfer to recipients that already carry identical or similar plasmids (Frost et al., 1994). These so called "entry exclusion" systems of conjugative plasmids can prevent attachment. Alternatively they may operate to prevent DNA transfer after Mpf formation (Achtman et al., 1977). Once contact is established the donor and recipient are brought together, maybe by process of pilus retraction, to form a mating aggregate (Lawley et al., 2003a). It is at this point that a signal may be evoked to the Dtr system of the plasmid (Lengeler et al., 1999) to initiate the second phase of conjugation; DNA processing and transfer to recipient.

The first step in DNA processing is strand and site specific DNA cleavage of the conjugative plasmid. This occurs at the *nic* site of the origin of transfer (*oriT*). This is controlled and implemented by relaxases that, together with the contribution of other accessory proteins form a nucleoprotein complex referred as a relaxosome (Zechner et al., 2000). The relaxosome and its interaction with the Mpf system is synchronized by a coupling protein (CP) that is required for conjugative DNA transfer to occur (Cabezon et al., 1997; Gomis-Ruth et al., 2002). In chimeric systems, constituting Mpf and Dtr from different plasmids, it is the CP that determines whether conjugative transfer occurs or not. The CP protein can be seen as opening the gate for transfer (Schroder and Lanka, 2005). CPs have also been identified in other Type IV protein secretion systems, although not in all (Gomis-Ruth et al., 2002). The second step in the Dtr process is the generation of a single stranded DNA copy of the plasmid. It is likely that this occurs when the relaxosome and dtr/Mpf systems are linked by the CP. Following nicking at the *oriT,* the 5' end of the DNA is covalently linked to the relaxase and elongation may occur at the 3' end that results in displacement synthesis by a rolling circle mechanism (Pansegrau and Lanka, 1996; Byrd and Matson, 1997). This ssDNA is then transferred via the Mpf system where, once in the recipient cell, complementary strand synthesis occurs by host encoded machinery. Once the relaxase recognises the reconstituted *nic* site it joins the ends together thus circularising the donor plasmid (Pansegrau and Lanka, 1996; Byrd and Matson, 1997).

So far a basic model for conjugation has been outlined for plasmids that are self transmissible as they encode the necessary Mpf and Dtr systems. However, plasmids are able to exploit transfer via conjugation by relying on other plasmids for Mpf and sometimes the CP and Dtr systems. Such plasmids are referred to as mobilisable plasmids. At the very least, mobilisable plasmids only have a compatible oriT, although many encode their own Dtr system (mob genes). Examples of well studied mobilisable plasmids are ColE1 (Warren et al., 1978) and RSF1010 (Willets and Crowther, 1981). Additionally, it must also be noted that the transfer of DNA in plasmid conjugation events is not always unidirectional from donor to recipient, but can be bidirectional. This phenomenon where plasmid transfer is accompanied with the acquisition of a non conjugative plasmid from the "recipient" is referred to as retrotransfer (Mergeay et al., 1987). Since its discovery it has been observed to be widespread, occurring in many different bacterial genera and by plasmids of many different incompatibility groups (Ronchel et al., 2000).

1.2.2. Plasmid persistence: A trade off between burden and benefit

Before plasmid diversity and adaptation in the natural environment is introduced (next section) it is worth considering why plasmids persist. Although, it is accepted that no single explanation can be given to encompass all plasmids.

It has long been established that plasmid carriage is not benign. A series of *in vitro* competition studies has demonstrated this (Examples include: Reinikainen and Virkajarvi, 1989; Lenski and Bouma, 1987; Sterkenburg et al., 1984). In the absence of positive selection plasmid containing hosts will be out-competed by their plasmid free counterparts, and the plasmid is fated to become cured from the population. The reason for this burden is, in part, explained by the extra metabolic cost associated with plasmid replication, but mostly by expression of plasmid encoded genes (Bentley et al., 1990). There are a range of burdens exerted by plasmids on their hosts dependent on their size, copy number and level of gene expression. In order for a plasmid to persist in the population it must overcome, or in someway circumvent, this burden. There are essentially two strategies, which are not mutually exclusive and better seen as the two extremes of a continuum. At the one extreme, plasmids are regarded as analogous to parasites, persisting while contributing absolutely no fitness benefit at any time to their respective host. At the other, plasmids overcome their burden by entering into a mutulistic or symbiotic relationship with their host.

In the analogy that plasmids are parasites it may be possible to draw parallels between plasmids and parasitic bacteriophages. Bacteriophage such as P1 have two developmental pathways. Firstly, the lytic pathway which results in the lysis of the host and dissemination of new phages. Secondly, a lysogenic pathway in which the phage exists as a prophage in the cell until triggered to enter the lytic pathway. Prophage are mostly integrated into the host chromosome, however, P1 lysogenises its host as an autonomously replicating low copy number "plasmid" (Lobocka et al., 2004). Like other lysosgenic phage, P1 in response to cell SOS signals is able to spread and persist using the viral option. It may be possible that plasmids can similarly spread, not by host lysis, but by using conjugative transfer mechanisms. The idea that plasmids could persist by transferring at sufficiently high rates to overcome the fitness cost associated with their burden has been investigated by a combination of experimental models and experimental data (Stewart and Levin, 1977; Levin et al., 1979; Simonsen et al., 1990). These have shown that plasmids may persist by parasitism under the correct cell densities. However, although theoretically possible in natural environments, the transfer rates and cell densities are probably insufficient to explain persistence by parasitism alone, and additionally not all plasmids are conjugative (Simonsen, 1991; Bergstrom et al., 2000). Nevertheless, there is some evidence that plasmids could at least periodically be maintained in environmental populations by parasitic means. For example, Bjorklof et al. (1995) demonstrated this experimentally by showing that the rate of transfer and loss of a marked pseudomonad plasmid was in equilibrium in the plant phytosphere.

If purely parasitic means do not explain plasmid persistence alone, are mutualistic or symbiotic relationships a more likely explanation for most plasmids? This mutualistic relationship could be explained by either (or both) of two means. Firstly, that the plasmid and host co-evolve to overcome the burden associated with plasmid carriage. Secondly, the plasmid encodes traits that are of infrequent benefit to the host in its natural habitat. The first of these two means has been demonstrated experimentally *in vitro* (examples include Modi and Adams, 1991 and Dahlberg and Chao, 2003). Coevolution by mutations and rearrangements in both plasmid and chromosome resulted in plasmid-host associations that were fitter than its plasmid free counterpart. In such an example the relationship can be described as mutualistic. However, coevolution itself is unlikely to describe plasmid persistence, as in such a scenario there would not be a requirement for plasmids to encode post segregation killing systems, which are found on some plasmids. The identification of these systems suggests the plasmid and host relationship, under certain conditions, to be antagonistic rather than mutualistic (Summers, 1996).

The second possible means of plasmid persistence is the carriage of beneficial traits. A general feature arising from the analysis of plasmid genomes is that they carry genes that encode proteins which do not have "housekeeping" functions, and are postulated to only be selected intermittently. This has led some to romantically suggest they act as a "lending library" (Summers, 1996). In this framework the burden associated with plasmid encoded genes is distributed among the population rather than on the individual, yet the genetic information is available to all species and habitats. Eberhard (1990) suggested reasons why plasmids contain these non-housekeeping genes. Firstly, plasmids are more likely to receive a benefit from traits that are not present on the chromosome i.e. functionally unique. Secondly, these traits are not restricted by the host reproduction rate since plasmids can transfer horizontally. Thirdly, plasmids can sample different genetic backgrounds simultaneously, some of which will be superior to the original host in different environmental conditions. It has been suggested that it is the carriage of genetic information that is only intermittently selected, and the ability to transfer to new hosts (sweeping through the population) that best explains plasmid persistence (Bergstrom et al., 2000; Turner et al., 1998). Indeed, the rapid explosion of plasmid carrying pseudomonads and subsequent decline over a sugar beet season may support this idea (Bailey et al., 2001). However, if a plasmid becomes established within a particular niche it may collect genes that are beneficial under similar conditions, (referred to as the local adaptation hypothesis) (Eberhard, 1990), as these genes will gain an advantage by being carried on the same plasmid. In this scenario plasmids can be seen as niche/habitat specialist (plasmid specialisation in the environment is considered further in section 1.2.3.1). Ultimately this collecting of locally relevant genes may go further and lead to a reduction in horizontal transfer, in favour of vertical transmission. Here the relationship is a symbiotic one and an example may be the megaplasmids of the Rhizobia, e.g. the pSymB of *Sinorhizobium meliloti* (Finan et al., 2001). It is at this point on the scale that the boundary of plasmid or secondary chromosome becomes less defined i.e. the element displays both chromosomal and plasmid like characteristics. For example, the pSymB plasmid cannot be cured from *S. meliloti* as it contains a small percentage of genes that are of "housekeeping" function suggesting chromosome-like characteristics. Conversely it still encodes a functional plasmid replication mechanism and a large percentage of genes which can be seen as accessory, suggesting plasmid-like characteristics. The issue of secondary chromosomes is considered further in chapter 5.

Based on the two extremes of plasmid existence by parasitism or symbiosis with host, in reality most plasmids probably exist to varying degrees between the two. Where a plasmid lies on this continuum is likely a reflection of the evolutionary history of the plasmid, diversity of bacterial habitats experienced and environmental fluctuations, suggesting plasmids are not statically fixed on this scale. Moreover, most studies of plasmid persistence rely on mathematical models that have been based on observations of small plasmids *in vitro*. These environments are largely homogenous. It is likely that in natural environments how plasmids persist is complex. In the next section the heterogeneous soil and plant environments are introduced.

1.2.2.1 The soil and the plant phytosphere environments

The soil environment is heterogeneous in terms of the distribution of liquid, gaseous or solid compounds, and this heterogeneity is dynamic, varying both over time and space (van Elsas et al., 2000; 2003). Studies have indicated that the nutrient availability in bulk soil is limited, and also that bacteria are found mainly adsorbed to surfaces, such as clay particles, in the form of microcolonies. As such they are physically restricted in their ability to interact with bacteria other than those locally (van Elsas et al., 2000). The lack of nutrients and physical separation consequently means bacterial populations in bulk soil are low in terms of density, growth and metabolic activity (van Elsas et al., 2000). Although it is not known which of these restrictions are responsible, plasmid transfer is low in bulk soil (Schwaner and Kroer, 2001; Lilley et al., 1994).

However, there are regions in the soil termed "hot spots" for microbial activity that enable increased growth, microbial mixing and horizontal gene transfer (Davison, 1999). These "hot spots" usually occur at the mineral organic interphases such as the surfaces of living or deceased higher organisms e.g. soil invertebrates, their waste products and the plant phytosphere constituting the rhizosphere and phyllosphere (van Elsas et al., 2003).

The rhizosphere can be simply defined as a "biologically active zone of the soil around plants roots that contains soil-borne microbes including bacteria and fungi" (Singh et al., 2004). The population structure and microbial interactions within this environment are complex and as yet not fully understood. For example, varying interactions exist between the microbe and plant i.e. symbiotic, pathogenic or associative. This inherent complexity is further exacerbated by the interactions between the microbes themselves e.g. competition and predation (Morgan et al., 2005).

The increased activity in the rhizosphere is dependent on both biotic and abiotic factors, but can be largely attributed to the increased concentration of nutrients in the rhizosphere compared to the bulk soil, due to the process of plant rhizodeposition. Rhizodeposition is the transfer of carbon and other nutrients to the soil surrounding the root. Rhizodeposits are caused by active secretion and passive diffusion of compounds (e.g simple sugars, enzymes) and also the release of compounds by the decay of sloughed root cells (e.g. lignin and cellulose) (Shaw and Burns, 2005). Like the rhizosphere the phyllosphere, encompassing the aerial parts of plants, represents a nutrient rich surface that supports complex microbial communities and activities (van Elsas et al., 2000). Nevertheless, survival in this environment can be harsh, as in this environment the microbes must tolerate rapidly fluctuating temperatures and humidity as well as exposure to ultraviolet radiation (Lindow and Brandl, 2003).

The soil and phytosphere represent highly complex environments with respect to both abiotic and biotic factors. In the next sections the influence environment has on shaping bacterial and plasmid genomes is considered.

1.2.3. Bacterial diversity, chromosome size and redundancy in response to habitat.

From the first 60 microbial genomes sequenced, Doolittle (2002) identified that chromosome size and number of coding sequences was correlated. These analyses also indicated that chromosome size was correlated to habitat; smaller chromosome sizes were associated with obligate intracellular pathogens, and larger ones are associated with free living bacteria that are only intermittently influenced, if at all, by higher organisms. For bacteria that form transient rather than obligatory relationships with higher organisms, such as the commensal bacteria of animal guts, intermittent genome sizes were observed. This is illustrated by the genome sizes of different Gamma proteobacteria. The obligate intracellular pathogens of the *Buchnera* sp that have chromosome sizes of less than 0.65 Mbp compared to the animal associated *Salmonella* sp of ~4.9 Mbp and the soil associated bacterium *Pseudomonas syringae* which has a chromosome size of 6.3Mbp (Source: www.ncbi.nlm.nih.gov/).

This difference in chromosome size can be explained by the heterogeneity of habitat. The intracellular environment is relatively static and homogenous compared to that of soil which, as already stated, is dynamic both spatially and temporally. In the intracellular environment, selection and competition within the niche are high. This results in chromosome size reduction (streamlining) and a reliance on host materials (Cole et al., 2001; Doolittle, 2002). Here the loss of genes can be seen as irreversible in terms of life style, with the bacterium becoming a specialist and, as such, niche restricted. On the other end of the scale the bacterium is a generalist encoding genes that can either enable the bacterium to tolerate or exploit a wider range of conditions. The intracellular bacterium *Mycobacterium leprae* offers evidence for this specialisation to niche. Analysis of the genome sequence has revealed that over half the chromosome is non-functioning and "decaying", consisting of pseudogenes. Moreover, in comparison with the genome of *Mycobacterium tuberculosis*, *M. leprae* shows a mosaic arrangement as a result of extensive gene deletion that has led to size reduction (Cole et al., 2001).

Larger chromosome size is not only a product of novel genetic functioning to cope with a dynamic environment, but also due to genetic redundancy. Such redundancy could be due to gene duplication or increased number of parallel biochemical pathways (Turner et al., 2002). For example, redundancy has been identified in the sequenced *Rhizobium etli* genome (Gonzalez et al., 2006), although this is not the first observation of redundancy in rhizobia (Galibert et al., 2001). Also, redundancy has been identified in *Burkholderia fungorum* (Marx et al., 2004) and in the genome of *Pseudomonas aerugionoas* PA01 (Chen et al., 2004). More recently the comparison of the proteomes of the 115 completed prokaryotic genomes, for orthologous groups using COGs (Clusters of Orthologous Groups), has revealed that larger chromosomes encode significantly more regulatory, transport and secondary metabolite genes than smaller chromosomes (Konstantinidis and Tiedje, 2004).

The explanations for genetic redundancy are at least two fold. Firstly, mutations that cause loss of function of an essential metabolic pathway are "backed up". Secondly, genetic drift among duplicated sequences and increased redundancy enable the genome to evolve novel functions that even if only subtly different could be advantageous. The correlation of redundancy with habitat has been described previously (Krakauer and Plotkin, 2002; Turner et al., 2002). Krakauer and Plotkin produced a series of mathematical models investigating the level of redundancy in response to population size and the distribution and level of selective pressures. In the eukaryotic intracellular environment selection is likely to be high and localised. Mutations accumulated in this environment are purged from the population as it is forced through a narrow genetically selective window (bottleneck) for survival. In this instance it can be said that the population is displaying antiredundancy. Conversely, in an heterogeneous environment such as soil, the distribution of selection is likely to be less intense, more widely distributed, dynamic and weakly counter selected for, thus redundancy is tolerated (Turner et al., 2002). In fact population diversity is increased and maintained in highly heterogeneous ecosystems where conditions are dynamic.

Large genome size, genetic redundancy, and a generalist ecological strategy are a feature of bacterial chromosomes from soil environments. The question is whether this is also true of plasmids that are identified in these environments?

1.2.3.1 Inferences for plasmids in the soil environment

Like bacterial diversity which is enormous in the soil environment, this diversity is paralleled in the plasmid community. Plasmids have frequently been isolated from soil and have been found in all bacterial communities to date (Sorensen et al., 2005). A number of studies have investigated plasmids associated with the plant phytosphere. Examples include Lilley et al. (1994); Sundin et al. (1994); Schneicker et al. (2001). In this environment plasmids have been associated with a myriad of host adaptive traits including the synthesis of plant growth regulators, heavy metal resistance, nitrogen fixation and U.V resistance (Bailey et al., 2001). At the genomic level it is possible to draw parallels between what is observed for bacterial chromosomes and plasmid genomes in different habitats. However, unlike chromosomes in the soil environment that display a generalist ecological strategy, there seems to be evidence to suggest plasmids, in contrast, are highly specialised to a particular niche/habitat (local adaptation hypothesis).

The first evidence of this comes from some of the large plasmids of the rhizobia. Rhizobia include the tumour inducing pathogens and nitrogen fixing symbionts of plants constituting the genera *Agrobacterium*, *Sinorhizobium*, *Mesorhizobium*, *Azorhizobium*, *Bradyrhizobium* and *Rhizobium* (Turner et al., 2002). The agrological importance of the rhizobia has led to numerous chromosomal and plasmid sequences being obtained, and their distribution and diversity in the natural environment has been extensively studied. To date six chromosomes and over 15 rhizobial plasmids have been completely sequenced and deposited in the public databases. Rhizobia are an extremely diverse group but a common observation is that most have one or more accessory plasmids that are usually of low copy and frequently unitary. These plasmids are considered to be derived from ancestral *repABC* replicon and tend to be large, ranging from 40 kb to over 2 Mb, the later representing the largest plasmid to be sequenced to date. The *repABC* plasmids of the rhizobia are genetically diverse, however there are a number that are functionally well characterised, such as the symbiotic associated pSymB plasmid of *S. meliloti* (Finan et al., 2001), pNGR234 of *Rhizobium sp.* NGR234 (Freiberg et al., 1997) and the pTiC58 plasmid of *A. tumefaciens*. Genomic analyses of these plasmids have shown that considerable numbers of the phenotypic traits encoded are related to phytosphere adaptation and survival. It has also been observed that these plasmids are restricted to hosts that form intimate relationships with plants, and are largely absent from those persisting freely in the soil (Segovia et al., 1991). These observations suggest that the rhizobial symbiotic and tumour inducing plasmids are highly specialised to the plant phytosphere.

The pSym and pTi rhizobial plasmids are only one example of plasmids that display specialisation to the phytosphere. Functionally well defined plasmids related to plant pathogenesis are found associated in other genera that are considerably smaller than the symbiotic plasmids of the rhizobia. For example, the pathogen associated plasmids of *Xylella* and *Xanthomonas* (1.3-65Kb) (Monteiro-Vitorello et al., 2005) and those of the *P. syringae* pPT23A plasmid classes (size 30-150 kb) (Bender and Cooksey, 1986; Sesma et al., 2000) are two further examples of plasmid specialisation in the phytosphere environment. Moreover, even for plasmid groups where functional information is not available there are indications towards a specialist strategy. For example, the temporal succession of plasmids over a sugar beet growing seasons (referred to in section 1.2.2) can be interpreted as specialisation.

However, although there seems to be considerable evidence to suggest plasmid specialisation, particularly from plasmids that are functionally well characterised, the issue may be less clear cut. Where as the pSym and pTiC58 plasmid represent a good example of a specialised strategy, other rhizobial plasmids may not. For example, the recently sequenced genome of *R. etli* includes the sequences of six indigenous plasmid replicons (Gonzalez et al., 2006). These plasmids are large (184-642kb), can be easily cured, show considerable amounts of functionally undetermined capacity and display functional redundancy. For example, paralogous genes (gene duplication), not orthologues (multiple genes of a common ancestral origin) have been identified, particularly of ABC transporters and genes involved in small molecule metabolism, suggesting functional redundancy (Gonzalez et al., 2006). Although large genome size and functional redundancy is synonymous with what is observed with chromosomes in the soil/phytosphere environment, it does not *per se* offer evidence of a generalist ecological strategy. However, the amount of uncharacterised genetic capacity may represent fitness to multiple different habitats, and therefore the possibility that these plasmids provide mixed traits and represent a more generalist strategy.

It is most likely that plasmids in the soil/phytosphere environment are highly specialised and encode genes that are locally adapted. However, there is a general paucity of sequencing and functional characterisation of the plasmid community from this environment. As a result the issue of whether plasmids are specialists cannot be resolved. However, it is possible that there is not a single unifying theory of plasmids in the soil environment. Obtaining sequence and functional information may clarify the situation. In the next section some of the benefits of obtaining sequencing information and the legacy of doing so are considered.

Section 1.3: The genomic era: why sequence genomes?

A genome by classic definition is the complete nucleotide complement of an organism. More recently the term genome has also been used to refer to the complete complement of an extrachromosomal replicon i.e. a plasmid genome refers to the complete plasmid sequence.

Historically, the majority of bacterial genomes that have been completely sequenced are those of clinically important pathogens, for example the lung disease pathogen *Haemophilus influenzae* Rd (Fleischmann et al., 1995) and the stomach disease pathogen *H. pylori* 26695 (Tomb et al., 1997). Additionally, genomes of bacteria that have phenotypes of agronomic importance have also been targeted e.g. the plant growth enhancing and biocontrol bacterium *Pseudomonas fluorescens* PF-5 (Paulsen et al., 2005), the plant disease causing *P. syringae* pv. tomato DC3000 (Buell et al., 2003) and the nitrogen fixing and tumour inducing rhizobia (see above section). Largely as a consequence of chromosome sequencing projects a number of plasmid sequences have been obtained. However, as with chromosomes, there have been targeted efforts towards sequencing particular plasmids, especially those of clinical or industrial importance. Examples include the clinically important IncPα and IncPβ Birmingham plasmids (Pansegrau et al., 1994; Thorsted et al., 1998) and the IncP-9 xenobiotic degradation plasmid, pWW0 (Greated et al., 2002).

The specific reasons why genomes are chosen for sequencing are manifold, but essentially the sequences have been obtained in an effort to better understand processes of disease in animals or plants, in the hope for control or to better understand nature's processes for commercial exploitation. Regardless of the specific aims of the project, obtaining complete sequence information is beneficial in two ways. Firstly, it enables *in silico* analysis and hypothesis generation and testing. Secondly, the sequence provides a genetic legacy in terms of a molecular "toolkit" with which to facilitate new wet experimental studies. These benefits should be viewed as complementary to each other with hypotheses generated *in silico* being tested in the laboratory, and vice versa.

1.3.1 *In silico* comparative analyses

The benefits of completing genome sequences and comparative analysis have already been illustrated in this introduction. For example, genome sequences have enabled the detection of recombination events, *IS* elements, other MGE, repeat motifs, gene duplications and genome reduction, which have given insights into the forces and mechanisms of bacterial evolution. Similarly this is also true for plasmid genomes, for example the sequencing of a number of related IncP and IncH plasmids have enabled their evolutionary history to be postulated (Dennis, 2005; Gilmour et al., 2004), a subject which is discussed further in chapter 3. Another benefit in the growth of sequence and functional data in the public databases is that putative phenotypic functions can be inferred to newly obtained sequences. These predictions enable a better understanding of the sequence in its ecological context as well as facilitating new computational and post-genomic studies. However, some of the most interesting findings of genomic sequencing are, firstly, that the degree of genetic diversity greatly exceeds what is expected, for example the genome of *E. coli* O157:H7 which strikingly contains over 1300 strain specific genes in comparison to *E. coli* K12 (Perna et al., 2001). Secondly, despite over 300 microbial genomes having been completely sequenced, in addition to billions of bp of partial bacterial sequences, averagely 20-40% of predicted CDS have no homology to other sequences in the databases or to only other putative CDS of unknown function (Source Genomemine; www.genomics.ceh.ac.uk/cgi-bin/genomemine/gminemenu.cgi). This situation has also been found in plasmid genome annotations (examples include, Freiberg et al., 1997; Kaneko et al., 2000). Large amounts of DNA of unknown function poses interesting questions i.e. is it simply non functional DNA or does it represent strain specific functions that are hitherto unseen? Faced which such questions it is clear that obtaining the sequence is only the first step in understanding the biology behind the sequences, which cannot be answered by *in silico* methods alone. These methods are complemented by comparative genomic array and global functional studies such as transcriptomics, proteomics and metabolomics. It is these whole genome investigations that are considered next.

1.3.2 The legacy of sequencing genomes

1.3.2.1 Hybridisation techniques

Immobilising DNA and challenging it with labelled specific sequences was first developed by Southern in the 1970's (Southern, 1975). The power of such a technique became readily apparent as a way of screening whole libraries for specific sequences e.g. plasmid bearing environmental isolates for specific incompatibility types (Kobayashi and Bailey, 1994). However, with the complete sequencing of genomes the technique has become popular in genomic analyses.

Recently, with decreasing sequencing costs there has been an increase in the sequencing of multiple strains of a species. Comparative analysis of closely related strains is desirable as the scope for analysis on genomes that are the only sequenced representative of a species or genus can be limited to the species or genus level, therefore missing the interesting strain specific information (Kim et al., 2002). Nevertheless, the legacy of a complete genome does enable the extraction of meaningful information from members of closely related groups where little or no sequence information is available via DNA arrays (comparative genomic hybridisations).

Additionally, the expression of the organism(s) represented on the array in response to different environments can also be studied using DNA arrays (transcriptomic hybridisations).

DNA arrays, whether for genomic comparison or transcriptomic studies, consist of multiple DNA probes. These can be synthetically produced oligonucleotides (i.e. affymetrix system, Lipshutz et al, 1999) or PCR amplified products that are immobilised in discrete locations on a solid support, traditionally a nylon membrane, glass or silicon. There are many variations in array design, construction and hybridisation, but essentially they all work on the same principle. The array is challenged with labelled test nucleotides that hybridise to complementary sequences (the immobilised probes) and this hybridisation is detected usually via radioactivity, chemiluminescence or fluorescence. Both genomic comparative and transcriptomic studies are considered below.

1.3.2.1.1. Comparative genomic hybridisation (CGH) experiments

The development of genomic arrays which represent a whole bacterial genome (usually at the gene level) as a technique to compare the genome composition of closely related species, strains or isolates was first developed for a pathogenic organism i.e. for *H. pylori* (Salama et al, 2000). The rationale of comparing the distribution of potentially thousands of genes globally, rather than on a gene by gene basis, has been revolutionary in identifying regions that could potentially be involved in pathogenesis i.e. present in pathogenic strains but absent in non-pathogenic relatives (Kim et al., 2002). Such studies of genome composition using genomic arrays have been coined as "genomotyping" (Lucchini et al., 2001). Recently, as the number of genomes sequenced has increased in the public database, multiple strains of the same organism have been applied to a single array, termed "pan-genomic arrays", such as the Pan-*Neisseria* genomic array (Snyder et al., 2004). However, as the technology has become more accessible genomic arrays have been constructed to investigate the genome composition of other elements including plasmid specific arrays e.g. a pPT23A array used to investigate the pPT23A plasmid family of *P. syringae* (Zhao et al., 2005).

1.3.2.1.2 Transcriptomic hybridisation experiments

The design of a DNA array, whether for genotypic investigations or for transcriptomic studies, is essentially the same. The difference is in the sequences that hybridised to that array. Genomotyping compares the sequence of two different genomic DNAs. Transcriptomics analysis on the other hand compares the total messenger RNA (mRNA) for an organism in response to environmental conditions and interactions it is experiencing, not only in the laboratory but also in natural settings (Xu, 2006). Transcriptomics produces a huge wealth of information and is a first step in functional characterisation of genomes revealed by sequencing projects. Its application has provided valuable insights into the mechanisms of virulence and other host organism interactions (Fraser-Liggett, 2005).

1.3.2.2. Protein expression experiments

Similar to transcriptomic studies, proteomics also investigates the variable expression of the organism's genetic capabilities. However, in contrast to transcriptomics that detects gene transcription (mRNA), proteomics is the identification of genes that are not only transcribed but also translated i.e. the proteins themselves. Therefore, proteomic investigations can be defined as a snap shot of the total expressed proteins (proteome) (Wilkins et al., 1996) of a cell population at a given time under defined physiological conditions (Guerreiro et al., 1999).

Proteomic investigations are essentially a combination of separating a complex mix of proteins based on size and pH by two-dimensional gel electrophoresis (2-DE analysis) (O'Farrell, 1975), and the subsequent identification of the separated protein spot by peptide fingerprints, and/or peptide sequences (Mass Spectroscopy), followed by genomic database searches (Langlois et al., 2003). As a functional technique proteomics has been used in a broad range of studies, including investigation of plant microbe interactions. For example, insights into how *Streptomyces coelicolor* colonises the small aquatic plant *Lemna minor*, have been gained by proteomic analysis (Langlois et al., 2003).

1.3.2.3. Other genomic investigations

Genomic investigations such as genomotyping and functional genomic studies, including microarray transcriptomics and proteomics, have become increasingly popular and useful in the post genomic era. However, the benefits of sequencing genomes or plasmid elements is not only restricted to these methods. Other genomic investigations include large scale mutation analysis and metabolomics, which have been employed to good effect in "mining" the biology from complete genomes. Moreover, non-genomic methods (i.e. focusing on particular genes or loci rather than the whole) including Q-PCR, targeted mutagenesis and targeted cloning, have also been useful in the elucidation of biochemical pathways and gene function, as well as of use in diagnostic and environmental monitoring.

1.3.3 Data handling: community level plasmid collection

At the individual project level comparative sequence analysis combined with both functional and genomic approaches will enable a better understanding of how bacteria, and indeed plasmids interact, persist and spread in that particular environment. However, how this relates to the biology of plasmids at the wider community level requires comparisons to that community i.e. comparisons to all previously sequenced plasmid genomes.

Over the last decade the number of complete bacterial genomes deposited in the public databases has increased exponentially, largely as a consequence of the reduction in the cost and recognition of the benefits of genomic sequencing (introduced above). As a whole these sequencing projects represent a considerable cost, largely to public funds. Nevertheless, they offer an opportunity for large scale comparative analysis to identify patterns not only between closely related strains or closely associated habitats but between bacteria at higher taxonomic and environmental extremes. An example of the power of community comparisons is the observations of chromosome size and bacterial specialisation (introduced above) (Doolittle, 2002). However, in order to facilitate such comparisons the bacterial chromosomal sequences and associated metadata (associated descriptive features of the sequence) must be stored and accessible for interrogation. For bacterial chromosomes that can be "pinned" on a taxonomic structure, the collection of genomes and their associated metadata is now quite well established (www.genomics.ceh.ac.uk/cgi-bin/genomemine/gminemenu.cgi). However, despite the recognition that plasmids play an integral role in bacterial evolution and adaptation, little attention has been given to collection and storage of all plasmid genomes making interrogation at the plasmid community level limited, and represents a waste of the information available to facilitate new computational studies.

Section 1.4 Aims of thesis

As discussed above, the diversity and importance of plasmids in host evolution and adaptation to the environment has been widely recognised. Many of these insights have been gained by plasmid and whole genome sequencing as well as subsequent genomic studies they afford. This has been achieved despite the absence of a coherent database in which to compare plasmids in the wider context of what has been sequenced previously (see section above). In this thesis, the aims are to use plasmid genomic studies to investigate a large phytosphere plasmid (introduced chapter 2) that has been demonstrated to impart fitness in natural settings. This plasmid is a representative of an ecologically well characterised collection of mercury resistant and self transmissible plasmids, isolated from natural phytosphere hosts at a single site. The first specific aim of this thesis is to obtain and annotate the complete plasmid sequence (representing the first plasmid genome to be sequenced from this collection), with the main intention of identifying putative determinants for the observed environmental fitness, as well as determine the molecular mechanisms of the phenotypes already described for the genome e.g. identify the mechanism of conjugative transfer, replication etc. Using the legacy afforded by the annotation and sequence, a second aim is to use comparative genomics studies to place this plasmid in the context of others isolated from the site in terms of gene content, and to investigate plasmid dynamics within the community. Building further on the legacy, a third aim is to employ expression techniques to investigate plasmid/host expression, *in vitro*, to differing environments. The final aim is to construct a database of all completely sequenced plasmid genomes and associated metadata (publicly available). By doing so, this will enable the plasmid to be positioned in context to all previously sequenced genomes and the current understanding of plasmid biology.

The aims can be summarised as follows

1) Obtain and annotate the complete nucleotide sequence of a large exogenously isolated phytosphere plasmid.
2) Asses the genetic diversity of plasmids isolated from the same environment.
3) Investigate the functional responses of the plasmid and its host to environments of different nutritional status.
4) Develop a plasmid genome database to enable the plasmid to be placed in the context of all completely sequenced plasmids.

These aims will enable the following hypothesis/questions to be addressed

In the introduction it was argued plasmids are environmental specialists that offer periodic benefit host. It is therefore hypothesised that the phytosphere associated plasmid will encode traits of adaptive benefit to that environment. The first question posed is can putative environmentally adaptive traits be ascribed to the plasmid genome? Further, can support for the ecological relevance of this plasmid be aided by identifying functional plasmid responses to differing environments simulated to represent conditions the plasmid may experience in nature?

The plasmid to be sequenced and analysed only represents one of a large collection of plasmids isolated from the phytosphere at a single geographical location. The third question posed is how representative is this plasmid of the local plasmid population and is there evidence of a common evolutionary origin?

As stated, past genomic projects have focussed on sequencing clinically important strains or strains of agronomic importance. Therefore, the fourth question posed is; how representative is the existing plasmid genome collection in relation the plasmid to be sequenced and other phytosphere plasmids?

Chapter 2: Plasmid pQBR103

2.1 Introduction

The phytosphere is a hot spot for genetic exchange and subsequent bacterial adaptation. What is clear from this environment is that conjugative plasmids play an integral, if not dominant role in this process (van Elsas et al., 2003). Since the 1990s, at a single site in Wytham Oxfordshire, numerous studies spanning several years have investigated the transfer proficient plasmids associated with the natural bacterial phytosphere community (examples include: Lilley et al., 1994; 1996, Lilley and Bailey, 1997a; Bailey et al., 2001). This extensive analysis makes these one of the best studied groups of environmental plasmids from a single geographical location to date.

Seminal investigations of the plasmids at this site were the isolation and analysis of naturally transferring plasmids from the sugar beet (*Beta vulgaris*) phytosphere (Lilley et al., 1994; 1996; Lilley and Bailey, 1997a). In these studies plasmids were captured by exogenous methods into pseudomonad, hosts or captured *in situ* by transferring into a marked pseudomonas strain colonising the phytosphere. In both methods mercuric (Hg^{2+}) selection was used, a common environmental plasmid encoded trait (Bailey et al., 1996). These plasmids were subsequently named the pQBR plasmids. The pQBR plasmids were large (100-500 kb) and based on RFLP analysis could grouped into five main genetically distinct types or groups (I-V). Three of these groups (I, III, IV) were found to be persistent at the field site with members being isolated over successive growing seasons (Lilley et al., 1996). In addition, members of these genetic groups, based on limited endogenous investigations, did not appear to be host specific or specific to the sugar beet phytosphere; Investigating fluorescent pseudomonads over a 0.75 km transect, containing a range of habitats (grazing pasture, woodland and open meadowland), pQBR types were isolated from a range of crop and wild plant species (including; maize, nettle, thistle and daisy), and a range of *pseudomonas* hosts (at least 15 based on REP-PCR analysis). It is likely, therefore, that these plasmids represent specialists to the phytosphere rather than specialists to any particular plant species or host.

Of the pQBR plasmid collection pQBR103 is by far the best characterised. pQBR103 was captured *in situ* in the sugar beet phyllosphere, by an indigenous *P. fluorescens* strain that was marked and reintroduced to sugar beet crops (Lilley and Bailey., 1997a). Based on

RFLP estimates, pQBR103 is large (~330 kb) and is a member of the group I plasmids, one of the most abundant plasmid groups identified in the sugar beet phytosphere (Lilley et al., 1996). Biological theory dictates a plasmid of this size is sure to exert a considerable metabolic burden on its host in the absence beneficial traits. Indeed this has been demonstrated experimentally by previous greenhouse and field studies. In these studies the effect of pQBR103 carriage on sugar beet colonisation fitness was determined by monitoring the fate of inocula over plant maturation (Lilley and Bailey, 1997b). In early growth stages it was found that pQBR103 imparts a relative reduction in fitness on *P. fluorescens* SBW25 in comparison to the plasmid free strain. However, on plant maturation plasmid carriage has no distinguishable effect on fitness. This demonstrates pQBR103 carries traits that are of temporal benefit in the phytosphere. What is unresolved is the function of these plasmid borne traits.

In vivo expression technology (IVET) (Osbourn et al., 1987; Mahan et al., 1993) was developed as a technique for identifying promoters that are induced in response to a particular environment (for review of the technique for the rhizosphere environment see Rainey, 1999). The environmental fitness phenotype is likely to be a combined effect of a number of disparate traits. Therefore, traditional methods of gene inactivation at a specific locus and subsequent screening for loss of fitness, is unlikely to be useful in investigating environmental fitness (Rainey, 1999). The benefit of IVET analysis is that it allows the detection of ecological genes on a genome wide scale by their positive contribution towards environmental fitness. Previously, IVET technology has been used as an approach to identify genes encoded by pQBR103 and *P. fluorescens* SBW25 that are specifically expressed in plant associated environments (Rainey, 1999; Zhang et al., 2004 a; b).

Based on IVET analysis, 37 pQBR103 gene fusions have been identified that are specifically transcribed in the phytosphere of sugar beet seedling (Zhang et al., 2004a). Size for size, this corresponds to over six times as many compared to the *P. fluorescens* SBW25 chromosome (Zhang et al., 2004b). This fits with the idea that the plasmid plays an integral role in adaptation to this environment. However, analysis of these plasmid IVET fusions has identified few environmental fitness traits, other than UV resistance (Zhang et al., 2004a; b). Combining the IVET analysis with phenotypes already known for the plasmid (Hg^{2+} resistance, Tra^{+}) it is still not possible to account for the ecological value of this plasmid. This makes it intriguing; therefore it was decided to completely sequence the plasmid genome, in the hope of affording a better understating of this enigmatic plasmid. In this chapter the focus is the elucidation and annotation of the pQBR103 genome sequence.

2.2 Chapter Aims

In this chapter the aims are

1) To extract the large pQBR103 plasmid DNA from the host chromosome at sufficient quantity, purity and quality for sequencing.
2) Determine the complete nucleotide sequence (To be completed by the Sanger Institute, Cambridge)
3) Annotate and analyse the pQBR103 genome.
4) Based on pQBR103 sequence information, develop a PCR screen to investigate diversity within the pQBR plasmid collection.

This will facilitate the following hypotheses/questions

1) It is proposed irrespective of isolation environment plasmids encode common maintenance strategies e.g. replication, copy number control, stability mechanisms (Thomas, 2000). Therefore, is it also true that pQBR103 encodes similar archetypal maintenance functions common to other plasmid genomes sequenced?

2) It is hypothesised that pQBR103 is a specialist of the plant phytosphere and as such encodes traits that are of adaptive benefit to its host. In addition to providing the genetic basis of known pQBR103 phenotypic traits (Inorganic mercury resistance and U.V resistance), what other ecologically relevant traits can be identified for pQBR103 to explain in its environmental fitness? More specifically, are these traits common to those ascribed to other environmental plasmids that have been sequenced (e.g. Chemotaxis, utilisation of plant growth regulators, utilisation of novel carbon sources), or are these traits novel and hitherto undescribed?

3) pQBR103 represents only a single plasmid from a large collection from a single geographical location. How does pQBR103 relate (using gross estimations) to other pQBR groups isolated over successive seasons (I, III, IV), is there evidence of intergroup genetic mixing and/or common ancestry?

2.3 Materials and methods

All chemicals and media used were from Sigma-Aldrich, UK and Oxoid, UK unless otherwise stated.

2.3.1 Bacterial strains used

The bacterial strains and plasmid that were used in this chapter are as listed in Table 2.1.

All strains were recovered from glycerol stocks by streaking onto PSA or PSA containing 14 $\mu g\ ml^{-1}$ $HgCl_2$ where appropriate, and incubated at 28 °C for 48 hours. From each plate a single colony was streaked onto PSA or PSA containing 27 $\mu g\ ml^{-1}$ $HgCl_2$ where appropriate, and incubated at 28 °C for 48 hours before use.

Bacterial strains and plasmids	Name	Plasmid Group[§]	Estimated plasmid size (kb)[§]	Reference
Plasmid free Pseudomonad strains	*P. putida* KT2440	-	-	Franklin et al., 1981
	P. putida UWC1	-	-	McClure et al., 1989
	P. fluorescens SBW25	-	-	Bailey et al., 1995
Plasmids hosted in *P. putida* UWC1	pQBR4	I	321	±
	pQBR41	I	301	±
	pQBR42	I	367	±
	pQBR44	I	130	±
	pQBR47	I	256	±
	pQBR29	III	174	±
	pQBR55	III	149	±
	pQBR57	IV	261	±
Plasmid hosted in *P. fluorescens* SBW25	pQBR103	I	330	Lilley and Bailey 1997a

Table 2.1. *Pseudomonas* stains and plasmids used in chapter 2. § Groups and size estimated by RFLP analysis described in Lilley et al. (1996). * pQBR103 was also used in a *P. putida* UWC1 host for the PCR survey and test for sensitivity to organic and inorganic mercury. ± plasmids from Lilley et al. (1996).

2.3.2 Extraction of plasmid and bacterial DNA using the CTAB method

All the bacterial strains and plasmids listed above (section 2.3.1) were grown to saturation by agitation overnight at 28 °C in 5 ml LB cultures under 27 μg ml^{-1} $HgCl_2$ selection, where appropriate. DNA was extracted by the CTAB method of Griffiths et al. (2000). Briefly, 0.5 ml of each culture was harvested and the supernatant removed. To each pellet 0.5 ml of CTAB extraction buffer was added (10% wt/vol CTAB in 0.7 M NaCl with 240 mM potassium phosphate buffer, pH 8.0) and 0.5 ml of phenol-chloroform-isoamyl alcohol (25:24:1), pH 8.0. Samples were added to Multimix 2 Matrix tubes (Bio-101, CA, USA) and lysed using a FastPrep FP120 bead beating system (Bio-101, CA, USA) for 30 s at a setting of 5.5 m/s. Tubes were then centrifuged at 4 °C for 5 minutes at 14,000 x *g*. The aqueous phase was removed and added to an equal volume of chloroform-isoamyl alcohol (24:1) and centrifuged at 4 °C for 5 minutes at 14,000 x *g*. The aqueous phase was removed and DNA precipitation was performed by the method outlined in Maniatis et al. 1982. Briefly 10% volume 3 M NaAc and 2 volumes 95% ice cold Et-OH was added followed by centrifugation at 4 °C for 30 minutes at 14,000 x *g*. The pellet was recovered and washed with 70% ice cold Et-OH and centrifuged for 20 minutes at 14,000 x *g* and dissolved in MillQ water. Extraction purity was determined by one dimensional agarose gel electrophoresis using Et-Br staining and U.V detection (Maniatis et al., 1982) (herein referred to as gel electrophoresis) and quantity was determined by Gene-Quant analyses (Amersham Pharmicia Biotecch)

2.3.3 Extraction and purification of pQBR103 plasmid DNA for sequencing

P. fluorescens SBW25 carrying pQBR103 was grown to saturation by agitation overnight at 28 °C in 50 ml LB cultures under 27 μg ml^{-1} $HgCl_2$ selection, and enriched using the sucrose gradient method of Wheatcroft and Williams (Wheatcroft and Williams, 1981). Briefly, Cells were harvested at 1,400 x g for 10 minutes and resuspended in 1.6 ml reagent A*, 0.6 ml of reagent B* was added and the mix frozen at -20 °C before being defrosted at room temperature. 2 ml of this mix was vigorously vortexed for 7 minutes on the top setting of a whirlimixer (Fisons,UK). 2 ml of the mix was transferred to a sucrose gradient$^+$ and centrifuged at 103, 745 x g (average) 141,371 x g (max) for 1 hour at 20 °C (Beckman centrifuge, J2-21 Rotor SW28). Fractions were taken from the gradient using a wide bore hypodermic needle. Then 15 µl of selected fractions were run on a gel (0.6% agarose gels, 1% TBE run overnight at 20 V). Gels were stained using Et-Br and visualised using UV detection. Samples which showed highest plasmid to chromosome DNA ratio were chosen for subsequent gel electrophoresis, extraction and recovery.

* Reagent A: 50 mM Tris pH 8.0, 50 mM EDTA, 5% v/v Dow Corning Antifoam RD Emulsion, 0.1 mg/ml Xylene Cyanol FF.

* Reagent B: 1 M NaOH, 1% SDS w/v

$^+$ Sucrose gradient: 20% w/v sucrose. 15 ml added to Beckman ultraclear 5/8" x 4" centrifuge tube. Frozen -20 °C, thawed, refrozen and thawed before use.

Further purification and enrichment of plasmid was achieved by running 50 µl of chosen fractions on 0.6% LMP agarose gel (1% TBE) at 60 V for 12 hours. Duplicate runs of samples were performed; one was post stained with Et-Br. The unstained replicate plasmid band was excised from the gel and recovered. An equal v/w of extraction buffer (0.5 mM EDTA, 50 mM Tris-HCL pH 7.8) was added to excised fragment and heated for 10 minutes at 65 °C. An equal volume of phenol (pH 8.0) was added and mixed before being centrifuged at 14,000 x *g* for 10 minutes and the top layer recovered, the phenol step was repeated a further time. DNA was Et-OH precipitated as described in section 2.3.2 above. An *Xba*I digest using the previously described method (Lilley et al., 1996) was used to confirm the extracted plasmid was pQBR103. After obtaining the completed pQBR103 sequence an artificial *Xba*I digest *in silico* was used to further confirm the plasmid sequence as pQBR103.

2.3.3.1 Estimation of plasmid enrichment

To determine the level of plasmid enrichment a 1-10^{-6} DNA dilution series was produced for both the CTAB *P. fluorescens* SBW25 carrying pQBR103 extraction and the enriched plasmid extraction. Each dilution was used as template for PCR amplification of a chromosomal sequence (16S rRNA gene) and two plasmid encoded genes (*merA* and *merR*). The disappearance of the PCR products, as visualised by gel electrophoresis, for 16S rRNA gene and *merA* and *merR* along the dilution series was recorded, and from this the ratio of plasmid to chromosome determined. Comparison of the ratios for the genomic and enriched plasmid extraction enabled the level of enrichment to be determined. At least a 100 fold enrichment was necessary for library construction and sequencing as this would equate to 1:4 of the sequencing library clones being of chromosomal origin by chance (based on the assumption that plasmid cell copy number was one and the sizes of the plasmid and *P. fluorescens* SBW25 chromosome were ~350 kb and ~6.5 Mbp respectively). Although, the exact ratio was never determined on completion of the plasmid genome sequence chromosome contamination was far lower, below 1:19).

The *merA* and *merR* primers used can be found in Table 2.2 the 16S DNA primers were (Given in 5' to 3' orientation) 530R (GTA TTA CCG CGG CTG CTG) and GC338F (CGC CCG CCG CGC CCC CGC CCC GGC CCG CCG CCC CCG CCC ACT CCT AGG GGA GGC AGC). PCRs were performed using BioLine (London, UK) reagents and the following conditions for 25 μl reactions: Template DNA 100 ng μl^{-1} (0.5 μl), Each primer 10 pMol stock (0.5 μl), 10 x PCR buffer (2.5 μl), DNTP 25 mM (0.5 μl), $MgCl_2$ 50 mM (0.75 μl) Taq polymerase 5 U μl^{-1} (0.5 μl), MilliQ water (19.25 μl). The cycling conditions were as follows Initial denaturation 3 min 94°C then (0.5 min 94 °C, 1 min 54 °C, 1 min 72 °C) 30 cycles. Final extension 10 min 72 °C, hold at 4 °C. Cycling reactions were conducted on a PTC-225 DNA engine tetrad (MJ Research, UK).

2.3.4 Library construction and sequencing

Library construction, sequencing and assembly were conducted by the Sanger Centre, Cambridge.

Purified pQBR103 DNA was randomly fragmented by sonication. Fragments in the size range of 2-4 kb, as determined by gel electrophoresis, were ligated into the *Sma*I site of pUC19. The ligated DNA was transformed into *E. coli* DH10B cells. The finished assembly was based on 4,508 paired end-reads from this library and 357 paired end reads from a smaller pUC19 library (inserts 1.4-2.0 kb). Collectively this gave an 8.64 fold coverage. Sequencing was by ABI BigDye V3.1 terminator chemistry on ABI3700 sequencing machines. The sequence reads were assembled using Phrap software (http://www.phrap.org/) and GAP4 (Bonfield et al., 1995), resulting in three contigs above

3 kb, all of which were attributed to pQBR103. Gaps were closed by oligo walking on bridging pUC clones using PCR.

2.3.5 pQBR103 annotation

2.3.5.1 Sanger annotation

First draft annotation was provided by the Sanger Institute Cambridge using manual annotation methods as previously described (Young et al., 2006)

2.3.5.2 Updating, modification and extending the annotation

The pQBR103 annotation was modified and updated by the author. Initially, open reading frames (orfs) above 100 amino acids were queried using the blastp program (Altschul et al., 2001) (version = 2.2.8 matrix = Blosum 62, word length = 3 gap penalties existence 11, extension, 1) against the NCBI non redundant database (www.ncbi.nlm.nih.gov/GenBank) (search date = 01/04/2004). CDS prediction was aided by Glimmer 2 and RBS finder software (Delcher et al., 1999; Suzek et al., 2001). Where there existed multiple potential CDS to choose from the decision was made manually based on genomic context, and homology searches of each orf. CDS predictions proved to be largely congruent with the CDS predicted by the Sanger institute. As findings did not differ significantly from the Sanger CDS calling ultimately the Sanger CDS predictions were used, with one exception. One CDS was identified to have a 10 tri-nucleotide repeat at the start of the CDS this was trimmed to the next available methionine (ATG) codon.

CDS similarity searches were performed independently by the author. Firstly, using big-blast (www.sanger.ac.uk/Software/ACT/Bigblast) (matrix = 1-3 existence = 5, extension = 2) pQBR103 was compared to the NCBI non redundant nucleotide database (www.ncbi.nlm.nih.gov/GenBank). Secondly, both Fasta3 (www.ebi.ac.uk/fasta/) (version = 3.47 matrix = Blosum 50, word length = 2, gap penalties existence -12, extension, -2) and Blastp (Altschul et al., 2001) (version = 2.2.12 matrix = Blosum 62, word length= 3, gap penalties existence 11, extension, 1) searches were used to compare the predicted proteome against the Uniprot database (www.ebi.uniprot.org/index.shtml) (date of search = 31/01/06). In addition, each CDS was assessed for Pfam matches (Finn et al., 2006) (date of search = 31/01/06). Genes were annotated based on their coverage and similarity to database homologues, taking into consideration their genomic context. Generally, consensus was found between the first draft annotation of the Sanger centre and independent similarity searching. However, approximately 40 CDS differed from the Sanger annotation. These changes represented updates, but also changes due to inaccurate annotation, identified by more thorough analysis of the search results, and where interpretation of the results was aided by the understating of the genomic context and plasmid biology. For example, a number of CDS were annotated based on their homology to previous pQBR103 sequences that have been deposited in the public database.

Repetitive elements were identified using, Reputer (Kurtz et al., 2001), MSATfinder (www.genomics.ceh.ac.uk/msafinder) and Emboss einverted, palindrome and etandem software (Rice et al., 2000) all using default parameters.

2.3.6 Phylogenetic analysis

Protein sequences for all phylogenies were downloaded from NCBI GenBank (www.ncbi.nih.gov). Accession numbers for each gene are given below in brackets. All sequences were aligned in ClustalX (V1.81) (Thompson et al., 1997) using the parameters stated for each phylogenetic alignment (see below). A Neighbor Joining method (Saitou and Nei, 1987) was used for tree construction excluding positions with gaps and corrected for multiple substitutions. Each tree was bootstrapped (1000 replicates) and values converted to percentages. All trees were viewed in Treeview (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

2.3.6.1 Phylogenetic analysis of the partitioning homologue ParA

In total, 18 proteins with homology to the ParA superfamilies including the pQBR103 putative ParA homologue (CDS 001) were aligned. The sequences for comparison to pQBR103 were chosen to represent a phylogenetically diverse range of known ParA superfamily proteins (Hayes, 2000) and those of close Blastp homology (Altschul et al., 1997) to the pQBR103 sequences (version = 2.2.12 Matrix = Blosum 62, word length= 3, gap penalties existence 11, extension, 1) (date of search = 20/05/2005). The sequences used and their designated protein names were *A. tumefaciens* pTiB6S3, RepA (A32812); *Alcaligenes eutrophus* MOL28, ParA (S60670); *Aeromonas punctatus* pFBAOT6, IncC

(YP067819); *Bacillus subtilis*, Soj (P37522); Bacteriophage P1, ParA (P07620); Bacteriophage P7, ParA (S06099); *E. coli* F plasmid, SopA (P62556); *E. coli* MinD (P18197); *E. coli* RK2, IncC (P07673); *Enterobacter aerogenes* R751, IncC (Q52312); *P. putida*, OrfB (AAC08068); *P. putida* pM3, ParA (AAD46128); *Pseudomonas alcaligenes* pRA2, ParA (AAD40334); *P. syringae* p4180A ParA (AAD50907); *S. meliloti* pMBA19a, IncC (AAX19280); *S. coelicolor*, Soj (AAC03484); *Rhizobium leguminosarum* pRL8JI, RepA (CAA61616). Alignments were made using the following parameters: opening penalty of 10, Gap extension of 0.20 using Blosum series protein weight matrix.

2.3.6.2 Phylogenetic analysis of the replication initiator protein RepA

In total 8 proteins including the two pQBR103 Rep protein homologues (CDS 0383 and CDS 0445) were aligned. The protein sequences for comparison to pQBR103 CDSs represented all the significant homologues in the non-redundant databases using Blastp (Altschul et al., 1997) (version = 2.2.12 matrix = Blosum 62, word length= 3, gap penalties existence 11, extension, 1) (date of search = 20/05/2005) with a significance value cut-off of E < 0.1. The following sequences were used (host species name, plasmid name, gene designation, if known, and accession number is given for each): *Buchnera aphidicola* pBPs2, RepAC (CAA10001.1); *Pseudoalteromonas haloplanktis,* RepA-like (CAI89597); *Acidithiobacillus ferooxidans* pTF5, Rep (NP_863597); *E. coli* RA1, RepA (CAA52023.1); *Marinobacter aquaeolei* (ZP_00817620.1); *Nitrosospira multiformis* (YP_413475.1). Alignments were made using the following parameters: Gap opening penalty of 10, Gap extension of 0.10 using Blosum series protein weight matrix.

2.3.7 Protein homologues: paralogue and xenologue analysis

Homologous pQBR103 proteins were compared and clustered using TribeMCL software (Enright et al., 2002), using inflation values of 5.0 4.0 3.0, 2.0 and 1.2. The inflation value influences the size of the protein clusters (granularity of the clustering). For “tight” clusters high values were used (4.0-5.0) broader clusters were detected using lower values (1.2-3.0).

2.3.8 Identifying candidate conjugative transfer regions

Two approaches were taken to identify candidate transfer regions in pQBR103. First the pQBR103 proteome was queried using Blastp program (Altschul et al., 2001) (version =

2.2.12 matrix = Blosum 62, word length= 3, gap penalties existence 11, extension, 1) against a custom blast database comprising of the following well characterised plasmid Dtr (responsible for DNA transfer and replication functions) and Type IV systems: IncF1 F factor Tra, (U01159); IncP R751 Trb/Tra (U67194); pTiC58 Trb/Tra (AE007871), as well as against the functionally characterised type IV export systems, not involved in plasmid transfer pTiC58 VirB/D, (AE007871); and *Bordetella pertusis* Ptl, (L10720), where brackets denote accession numbers. pQBR103 CDS with significant homology to these systems were recorded (below an expected value of 0.01 to avoid the potential of hitting non-homologous sequences). Secondly, the pQBR103 proteome was queried using the blastp program (Altschul et al., 2001) (version = 2.2.12 matrix = Blosum 62, word length= 3, gap penalties existence 11, extension, 1) against the Uniprot database (date of search = 31/01/06) and any region of contiguous CDS (within 500 bp of each other i.e. putatively co-transribed) containing any homology to characterised or putative pilus/Type III/IV homologues was recorded.

2.3.9 Mapping *in vivo* expressed sequences

IVET analysis was used previously to identify promoters in pQBR103 that were induced in the phytosphere of immature sugar beet seedlings (Zhang et al., 2004 a; b; Rainey, 1999). Publicly available IVET sequences from these studies were downloaded from the *P. fluorescens* SBW25 encyclopaedia (www.plants.ox.ac.uk/SBW25). Each IVET clone consisted of two sequences, the first the *bla* end sequence which indicates the insertion point of the IVET into the plasmid genome; and secondly the *dap* end sequence, which likewise indicates the insertion point, but also represents the point through which transcription was expected to be driven. The *dap* ends were mapped to the plasmid genome for the IVET clones where the *dap* end and *bla* sequence were of an expected distance from each other (~3 kb) (n=22), or where only a dap sequence was available for that clone (n=6). IVETs where the distance between the *dap* and *bla* sequence was larger than expected were excluded (n=8, size range 9-100kb). CDS that are potentially transcribed by the inducible promoters were defined as contiguous CDS (on the same strand and within 500 bp of each other) that were in the same orientation of the *dap* gene.

2.3.10 PCR survey of the pQBR plasmids

Two PCR surveys were performed to investigate the genetic diversity between different pQBR plasmids listed in Table 2.1. These plasmids represent the three pQBR plasmid types that were isolated over successive growing seasons (pQBR types I, III and IV) (Lilley et al., 1996). As pQBR103 was identified as a group I plasmid a number of group I plasmids of different sizes were chosen for comparison. The first PCR survey was designed to give an estimation of global similarity to the pQBR103 genome. The second survey was designed to target two particular regions of the pQBR103 genome (two regions of transmebrane/pilus/Type IV homologues). In both surveys primers were designed using primer3 software (Rozen and Staletsky, 2000) to amplify a product of approximately 250 bp.

The first PCR survey was designed on a partial pQBR103 genome sequence. Constituting pQBR103 IVET fusions that had been sequenced, and contigs from an unsuccessful in house sequencing project (~ 300 kb in over 100 contigs). To ensure maximum global coverage of the genome only a single primer pair was designed per contig or IVET fusion. In Table 2.2a the 19 primer pairs designed to amplify the regions of pQBR103 are shown. In total 15 primer pairs amplified within CDS, the other four amplified within intergenic regions. In the second survey (Table 2.2b) the complete pQBR103 sequence was available. For the two regions targeted, primers were designed for each CDS (16 in total).

Using the CTAB extraction method and PCR conditions above (sections 2.3.2 and 2.3.3.1), for both PCR surveys each designed primer pair was tested to be specific by amplifying an appropriately sized product from pQBR103, but not from any of the plasmid free strains listed in Table 2.1 (detected by gel electrophoresis). For the first survey all primers were further demonstrated to be specific by end sequencing using ABI BigDye sequencing technology (Applied Biosytems, UK).

For both PCR surveys, extraction of test DNA (listed in Table 2.1$^{\pm}$) and PCR conditions were as described above. Each survey was repeated twice, using appropriate positive and negative controls. Positive amplifications were recorded if an appropriate sized product was amplified from the test DNA. Products amplified from the pQBR plasmids by primers CDS 431 (*merA*), 435 (*merR*), 055, 371 and int0023 (Table 2.2a) were sequenced using

ABI BigDye technology (Applied Biosytems, UK). Sequences were compared using Sequencher software v4.2 (Gene codes corporation, USA)

± pQBR103 used in the PCR survey was in a *P. putida* UWC1 host. Plasmid pQBR57 was not included in the first PCR survey.

CDS	Annotated function	Primer Sequence	Start..Stop (complement)	Size (bp)
055*	DNA Helicase, HelA	55F:CTCGATCCAGCACCAAAC 55R:GCCGCTCGCTACTGTTAC	(60274..60291) 60043..60060	249
062	Conserved Hypothetical	62F:AGATGTGCCGGTGTCAGT 62RAGAACGCGAAGGAAACCT	69754..69771 (69987..70004)	251
144	Orphan	144F:CTCGACGTGCAGAACAAC 144R:GCCTGTATCGGTGGACAT	154241..154258 (154468..154485)	245
151	Nucleoid-associated protein	151F:TGAATGACGCCTACAACG 151R:AGCGCAATGCTGAGGTAG	(161266..161283) 161040..161057	245
156	Orphan	156F:TGTGCCATGTTCACAAGC 156R:GTTGTGCTCCACGCTAGG	166026..166043 (166257..166274)	250
249	Rrestriction enzyme related protein	249F:CCCAATACGCTGGTTACG 249R:AGCGTACAGGTGCGTAGG	246571..246588 (246804..246821)	253
286	Conserved hypothetical	286F:CCACGCTTGATCAGGAAC 286R:CGGCATTCGTGAAAATGA	271130..271147 (271358..271375)	246
297	Orphan	297F:TCCCCCTGAGAAGGTAGG 297R:GCAGTGACAGGTCGCAGT	(278438..278449) 278202..278219	254
310	Orphan	310F:TTCAAGCCTGAGCATTCC 310R:GGTAGTCCCCGCTGATTT	288942..288959 (289180..289197)	247
318	Oligoribonuclease. Orn	318F:GTGAAGTTCCGGCATCTG 318R:TGGCCAGTACCTCAATCG	(294609..294626) 294370..294387	257
361	DNA helicase, HelB	361F:TCAGGTGCGAGCATTACA 361R:TGTTCAACAAAGCGTTCG	(322046..322063) 321811..321829	251
371*	DNA helicase	371F:GCAGCACAACACTGCATC 371R:CGCATGTTTCTGGAGGAC	(330482..330499) 330259..330276	242
391	Conserved Hypothetical	391F:TATGAGGGCCATTCCAGA 391R:CGGTTAATCCCAGCACAG	(346462..346479) 346227..346244	253
431*	Mercuric ion reductase (MerA)	431F:GCGACCAGCTTGATGAAC 431R:CCGCAGTTCGTCTACGTC	378826..378843 379059..379076	251
435*	Mercury operon regulator (MerR)	435F:GAACCGGACAAGCCCTAC 435R:CGCAACAGTCATCAACCA	381594..381611 381829..381846	253
	Intergenic sequences			
Int023*	Between CDS 053 and 054 overlapping with 054	Int23F:CCTGAGTCCCAGTGCTGT Int23R:TCATTTCCAACGGAATGC	56739..56756 (56965..56982)	245
Int730	Between CDS 215 and 216 overlapping with 216	Int730R:ACACCATCAGCCTCAACG Int730R:CAACATGGACCCCATCAT	221220..221237 (221450..221467)	248
Int036	Between CDS 282 and 283 overlapping with 283	Int36F:CCGTACCGGAGTGGTTGT Int36R:TCTACGCCCCAGTATTCG	269294..269311 (269519..269536)	246
Int034	Between CDS 453 and 454 overlapping with 454	Int34F:CCCCGGAAAGTGATGTAG Int34R:GGGTCATCGTCGGAGACT	394592..394609 (394836..394853)	245

Table 2.2a, PCR primers designed from a partial pQBR103 sequence to survey pQBR plasmid diversity. Primer sequences are given in 5' to 3' orientation. * Products amplified from the pQBR plasmids (listed in Table 2.1) using these primers were sequenced and compared.

CDS	Annotated function	Primer sequence	Start..Stop	Size (bp)
27	Orphan	27F: CCGACAAGCTCAGCAACA 27R: CTTCACACGCCAGCCATT	24799..24816 25040..25057	259
28	type II and type III secretion system pilus Protein	28F: CTATGATTGGCGGCTTGG 28R: TTCCCGCTGACCGTAGAA	26021..26038 26253..26270	250
29	Orphan	29F: GGCCTCTGGCAGCTCTTA 29R: GCGTCGATGGAAGGATCA	27132..27149 27361..27378	247
30	Orphan	30F: TGACGATGTCGCTGAGTG 30R: TTGAGCCGTCCAGTTTCG	27962..27979 28195..28212	251
31	Orphan	31F: CCGAGTTTCCGAGGTCAA 31R: CCGCGAACTTCTTGGAGA	28606..28623 28831..28848	243
32	Pilus assembly-related protein	32F: AGCGGTTGCAAGGTTGAG 32R: TTCCTGCGCATGGTGTTA	30355..30372 30585..30602	248
33	Pilus biogenesis-like protein	33F: ACCTGCTCCAGCACTTCG 33R: CGGTAGGCAACAGCGATT	31317..31334 31538..31555	239
34	Transmembrane secretion-related protein	34F: CGCGAGCAAGCCATTACT 34R: TATTCCTTGGGCGTTTCG	31656..31673 31892..31909	254
35	Transmembrane pilus-related protein	35F: CGGGCAAGCTCAGATGAT 35R: TCGCTCTGCAGGAAAGGT	32884..32901 33121..33138	255
36	Transmembrane secretion-related protein	36F: TCAACGAAAGGCGTGTCA 36R: CCCAGTCCGCAATGATGT	33382..33399 33613..33630	249
470	CheB like methylesterase	470F: GGCACGCTTGAAGGATGT 470R: ATCCATTCCCATGCCTGA	411349..411366 411117..411134	250
471	Two-component regulator sensor histidine kinase fused response regulator protein	471F: ATAGCGCTGGCATCCTTG 471R: TGTCGCGCAGTTCGATAA	417289..417306 417058..417075	249
472	Chemotaxis signalling protein, CheR like methytransferase	472F: CATTGCTTCCGCGCTACT 472R: TGTTCGGATTCGCCAGTT	418123..418140 417896..417913	245
473	Methyl-accepting chemotaxis transducer Protein	473F: CGAACCAGCTCCACCAGT 473R: TCTTCCAGCGACGACACA	420189..420206 419952..419969	255
474	Chemotaxis signal transduction protein CheW	474F: AGGTGCCGAAGTTGCTCA 474R: CTGATCGCAATCCCGAAG	420875..420892 420642..420659	251
475	Two-component response regulator CheY like domain	475F: AGGTTCACGTCGCTGAGG 475R: GCCCGTGTGAAGGGTTTA	421260..421277 421031..421048	247

Table 2.2b PCR primers designed from a complete nucleotide sequence to target two particular pQBR103 genomic regions. Oligonucleotides are given in 5' to 3' orientation

2.3.11 Test of organic and inorganic mercury sensitivity

All 9 pQBR plasmids in *P. putida* UWC1 hosts (as listed in Table 2.1) and appropriate controls, *P. putida* UWC1 and *P. fluorescens* SBW25, were tested for sensitivity to inorganic ($HgCl_2$) and organic (phenylmercuric acetate (PMA)) mercury. All strains were spread plated onto PSA media, with 27 μg ml^{-1} $HgCl_2$ where appropriate, and incubated for 48 hours at 28 °C. Cells were recovered from the plates by adding 2 ml PBS. Each cell suspension was centrifuged at 14,000 x *g* at 4 °C for 2 minutes. Supernatant was removed and sufficient PBS was added to resuspend cells to a density of 10^8 ml^{-1}. A dilution series of 1-10^{-7} was constructed for each test strain,10 μl of which was spotted onto 8 PSA plates of no mercury and varying $HgCl_2$ and PMA concentrations ($HgCl_2$: 27.2, 5.44 and 2.72 μg ml^{-1}, PMA: 10, 5, 2, 0.2 μg ml^{-1}). Plates were incubated for 48 hours at 28 °C and growth at the highest dilution was recorded for each condition.

2.4 Results and Discussion

2.4.1 General overview of the pQBR103 sequence

The elucidation of the pQBR103 sequence revealed the plasmid to be 425,094 bp in size. The plasmid sequence and annotation is available under the accession number AM235768. The plasmid genome was predicted to be circular and had an average G+C content of 53.2%, lower than *P. fluorescens* SBW25 in which the plasmid was captured (60.5%) (www.sanger.ac.uk/Projects/P_fluorescens/). The plasmid was predicted to be 83.4% coding, represented by 478 CDS of an average size of 246 amino acids, giving a gene density of 1.124 CDS kb^{-1}. Only one CDS was annotated as a pseudogene. Comparing strands, there was a clear bias with the majority of CDS assigned to the forward strand, 357 (76%) compared to 121 (24%) on the reverse strand. Of the 478 CDS, only 94 predicted proteins could be ascribed a putative function (these CDS are listed in Table 2.3, overleaf). A further 101 had homology to other known or putative proteins of unknown function in the public databases (Uniprot). However, the vast majority of CDS, 283 (59%), encoded putative proteins that had no significant homologues in the public databases, here after these CDS are referred to as orphan genes (Siew and Fischer, 2003). In Table 2.4 the characteristics of the orphan, conserved and CDS ascribed a putative function is given. Of the 94 CDS with a putative function 74 (79%) had a significant blastp hit (expected value below $1e^{-3}$ chosen as cut-off to avoid hits to non homologous sequences) to a Pseudomonad chromosomal or plasmid sequence in the Uniprot database. For 37 of these CDS (39% of the 94 functional ascribed genes) the pseudomonad sequence was the most homologous (Table 2.3). This may suggest that pQBR103 has a natural or favoured relationship with pseudomonas hosts.

CDS Group	Number	Proportion of total	Protein size, amino acids (s.d)*
Functional	94	20%	344.9 (253.51)
Conserved hypothetical	101	21%	294.65 (294.65)
Orphan	283	59%	196.24 (124.79)
Total	478	100%	246.48 (190.52)

Table 2.4. Comparison of pQBR103s annotated proteins with and without homologues in the public databases. * Standard deviation.

In figure 2.1 the general features of pQBR103 is shown. This figure demonstrates the distribution of CDSs around the genome in particular the coding strand bias and distribution of orphan genes (coloured light green in circles 2 and 3). The region of the genome that may be involved in plasmid maintenance/core functions (i.e. plasmid conjugation, replication and partitioning) or has been ascribed a putative environmental function (i.e. organic/inorganic mercury, chemotaxis, UV resistance) is shown in circle 4. What is noteworthy is the maintenance/core functions are not tightly clustered into a particular region of the genome as may have been expected (Thomas, 2000) (more discussion of this is given in chapter 3). Another observation is that potential promoters induced in the sugar beet phytosphere (circle 5) (Zhang et al., a;b) were found not to be evenly distributed around pQBR103 but clustered to particular regions (discussed further in section 2.4.3).

		Closest Blastp homology to a either a psedomonad plasmid or chromosome sequence		Overall closest Blastp homology			
CDS	Annotation*	Species (Plasmid)	E-value	Taxa, Species (Plasmid)	E-value	% ID	Accession
1	Plasmid partitioning protein, ParA	*P. putida* (pWW0)	$1.0\ e^{-23}$	Alphaproteobacteria; *Rhizobium meliloti* (pMBA19a)	$4.0\ e^{-33}$	34.78	Q5BTN9_RHIME
2	Plasmid partitioning protein, ParB	*P. aeruginosa*	$1.0\ e^{-05}$	Gammaproteobacteria; *Xylella fastidiosa*	$2.0\ e^{-13}$	26.81	Q9PH83_XYLFA
28	Type II/III secretion system pilus protein, PilN-like homologue	*P. syringae* B728a	$1.0\ e^{-15}$	Deltaproteobacteria; *Desulfotalea psychrophila*	$3.0\ e^{-21}$	24.31	Q6AS33_DESPS
32	Pilus assembly-related protein, PilB/TapB-like homologue	*P. aeruginosa*	$1.0\ e^{-57}$	Firmicutes; *Thermoanaerobacter tengcongensis*	$6.0\ e^{-66}$	40.25	Q8RAG1_THETN
33	Type IV pilus biogenesis-like protein	No significant hit		Low homology	6.6	31.25	Q1QKR2_NITHA
34	Type II/IV transmembrane secretion-related protein	No significant hit		Chlamydiae; *Parachlamydia* sp. UWE25	$5.0\ e^{-05}$	22.49	Q6M9X9_PARUW
35	Transmembrane pilus-related protein, PilA/ComP-like homologue	No significant hit		Low homology	2.2	28.28	Q3SEB6_PARTE
36	General secretion pathway protein	*P. fluorescens* SBW25	$5.0\ e^{-05}$	Betaproteobacteria; *Burkholderia ambifaria* AMMD	$3.0\ e^{-7}$	38.96	Q3FDY7_9BURK
37	Arylsulfatase-activating protein-like homologue	No significant hit		Euryarchaeota; *Methanosaeta thermophila* PT	$3.0\ e^{-20}$	24.93	Q2CLU4_9EURY
38	Arylsulfatase-activating protein-like homologue	*P. fluorescens* PfO-1	$9.0\ e^{-4}$	Euryarchaeota; *Methanosaeta thermophila* PT	$3.0\ e^{-16}$	26.29	Q2CLU4_9EURY
51	Arylsulfatase-activating protein-like homologue	No significant hit		Gammaproteobacteria; *Vibrio parahaemolyticus*	$7.0\ e^{-6}$	24.52	Q87IA8_VIBPA
52	Twitching motility protein, PilT-like homolgue	*P. putida* KT2440	$4.0\ e^{-19}$	Deinococcus-Thermus; *Thermus thermophilus* HB8	$2.0\ e^{-22}$	34.82	Q5SHF6_THET8
53	GGDEF-family domain protein	*P. putida* F1	$5.0\ e^{-17}$	Alphaproteobacteria; *Agrobacterium tumefaciens* C58	$1.0\ e^{-22}$	29.96	Q8UB10_AGRT5
55	Plant-inducible DNA helicase, HelA	*P. fluorescens* Pf-5	$2.0\ e^{-13}$	Deltaproteobacteria; *Pelobacter carbinolicus* DSM 2380	$2.0\ e^{-145}$	33.19	Q3A3V6_PELCD
56	Catabolite gene activator family protein, Crp/Vfr-like homologue	No significant hit		Gammaproteobacteria; *Photorhabdus luminescens*	$2.0\ e^{-4}$	23.76	Q7MB98_PHOLL

Table continued overleaf

57	Ribosomal protein, RpsJ/NusE/S10-like homologue	No significant hit		Actinobacteria; *Rubrobacter xylanophilus* DSM 9941	6.0 e^{-5}	32.65	Q3X1J8_9ACTN
65	Deoxyribonuclease	*P. putida* KT2440	1.0 e^{-12}	Gammaproteobacteria; *Nitrosococcus oceani* ATCC 19707	2.0 e^{-44}	41.45	Q3JF75_NITOC
73	Catabolite gene activator family protein, Crp/Vfr-like homologue	*P. putida* KT2440	5.0 e^{-26}	Gammaproteobacteria; Marine proteobacterium HTCC2207	1.0 e^{-26}	31.25	Q1YVH0_9GAMM
74	Plant-inducible DNA helicase, HelC	Top hit		Gammaproteobacteria; *P. syringae* 1448A	5.0 e^{-117}	33.92	Q48IC3_PSE14
76	Restriction-modification methylase	No significant hit		Betaproteobacteria; *Ralstonia eutropha* (pHG1)	5.0 e^{-104}	48.27	Q7WX17_RALEU
80	Transmembrane rhomboid family protein	Top hit		Gammaproteobacteria; *P. syringae* 1448A	2.0 e^{-68}	65.93	Q48FU5_PSE14
84	Glutathionylspermidine synthase	Top hit		Gammaproteobacteria; *P. syringae* 1448A	2.0 e^{-178}	74.29	Q48FU9_PSE14
97	DNA-binding domain protein	*P. putida* (pWW0)	4.0 e^{-10}	Gammaproteobacteria; *Shewanella denitrificans* OS217	2.0 e^{-19}	28.17	Q3NZV3_9GAMM
103	Response regulator domain protein	*P. fluorescens* Pf-5	5.0 e^{-08}	Bacteroidetes; *Salinibacter ruber* DSM 13855	4.0 e^{-11}	38.39	Q2RZC9_SALRD
104	Type IV leader peptide processing enzyme	Top hit		Gammaproteobacteria; *P. aeruginosa*	6.0 e^{-75}	48.96	LEP4_PSEAE
105	Site-specific recombinase, Integrase family, Int	Top hit		Gammaproteobacteria; *P. syringae* DC300	1.0 e^{-85}	42.82	Q881N3_PSESM
110	Site-specific recombinase, Integrase family, Int	*Pseudomonas* sp. ND6 (pND6-1)	6.0 e^{-16}	Gammaproteobacteria; *Xanthomonas campestris* 85-10 (pXCV183)	3.0 e^{-86}	45.19	Q3C033_XANCS
113	Plasmid partitioning protein, ParB	*P. putida* F1	6.0 e^{-10}	Actinobacteria; *Symbiobacterium thermophilum*	7.0 e^{-11}	43.06	Q67J37_SYMTH
119	Transmembrane thiol:disulfide interchange protein, DsbD-like	*P. syringae* DC300	2.0 e^{-75}	Gammaproteobacteria; *Azotobacter vinelandii* AvOP	1.0 e^{-81}	36.93	Q4IWJ7_AZOVI
123	Transmembrane protein	Top hit		Gammaproteobacteria; *P. resinovorans* (pCAR1)	8.0 e^{-29}	40.23	Q8GHV0_PSERE
124	Zn-dependent protease with chaperone function	Top hit		Gammaproteobacteria; *P. aeruginosa*	4.0 e^{-63}	54.65	Q9HVF9_PSEAE
126	Transmembrane protein, TolA-like homolgue	Top hit		Gammaproteobacteria; *P. aeruginosa*	1.0 e^{-13}	29.65	TOLA_PSEAE

Table continued overleaf

128	DNA-binding domain protein	*P. syringae* B728a	2.0 e^{-4}	Betaproteobacteria; *Polaromonas* sp. JS666	2.0 e^{-36}	46.53	Q4ASP6_9BURK
131	Transmembrane thiol:disulfide interchange protein, DsbD-like	Top hit		Gammaproteobacteria; *P. syringae* DC300	6.0 e^{-8}	27.61	Q87VS7_PSESM
134	Transmembrane autotransporter	Top hit		Gammaproteobacteria; *P. syringae* B728a	6.0 e^{-129}	40.65	Q4ZYT2_PSEU2
149	NAD-dependent deacetylase	Top hit		Gammaproteobacteria; *P. fluorescens* Pf-5	1.0 e^{-71}	57.02	Q4KDX3_PSEF5
151	Nucleoid-associated protein, NdpA-like homologue	Top hit		Gammaproteobacteria; *P. fluorescens* Pf-5	1.0 e^{-154}	81.08	Q4KHU2_PSEF5
155	Pilin-related protein, PilV-like homologue	*P. aeruginosa* (pKLC102)	5.0 e^{-05}	Gammaproteobacteria; *Serratia entomophila*	1.0 e^{-6}	33.33	Q7BQX0_9ENTR
157	UV resistance protein, RulA	Top hit		Gammaproteobacteria; *P. putida* (pNAH7)	7.0 e^{-34}	53.9	Q1XGP5_PSEPU
158	UV resistance protein, RulB	Top hit		Gammaproteobacteria; *P. putida* (pNAH7)	1.0 e^{-149}	59.86	Q1XGP6_PSEPU
160	Conjugal transfer protein, TrbN-like homologue	No significant hit		Low homology	3.30 e^{-1}	33.33	Q47073_ECOLI
171	Ribonuclease HII	Top hit		Gammaproteobacteria; *P. fluorescens* PfO-1	2.0 e^{-75}	72.87	RNH2_PSEPF
175	Transcriptional regulator DnaK suppressor protein, TraR/DksA-like homologue	Top hit		Gammaproteobacteria; *P. syringae* DC300	2.0 e^{-19}	44.09	Q87ZE2_PSESM
178	DNA-binding protein, Hu	*P. aeruginosa*	2.0 e^{-25}	Gammaproteobacteria; *Methylococcus capsulatus*	2.0 e^{-25}	61.8	Q60BE5_METCA
181	Exodeoxyribonuclease, RecD/TraA-like homologue	No significant hit		Alphaproteobacteria; *Acidiphilium cryptum* JF-5	1.0 e^{-18}	20.94	Q2DE92_ACICY
182	Conjugal transfer TraG-family coupling protein	*P. syringae* DC300 (pDC300A)	7.0 e^{-12}	Betaproteobacteria; *Burkholderia vietnamiensis* G4	3.0 e^{-33}	25.68	Q4BGZ3_BURVI
188	Conjugal transfer TraB-family topoisomerase	No significant hit		Low homology	4.30 e^{-2}	21.84	Q54WI5_DICDI
189	Conjugal transfer assembly protein, TraV-like homologue	No significant hit		Gammaproteobacteria; *Legionella pneumophila* Philadelphia 1	6.0 e^{-5}	33.64	Q5ZTS3_LEGPH

Table continued overleaf

191	Conjugal transfer TraC-like homologue	*P. resinovorans* (pCAR1)	2.1 e-^{18}	Gammaproteobacteria *Proteus vulgaris* (Rts1)	2.0 e^{-29}	30	Q8L274_PROVU
209	DNA primase, DnaG-like homologue	No significant hit		Spirochaetes; *Borrelia burgdorferi*	8.0 e^{-8}	23.43	PRIM_BORBU
213	GGDEF two-component response regulator	Top hit		Gammaproteobacteria; *P. aeruginosa*	3.0 e^{-58}	38.06	Q9HUW7_PSEAE
249	Restriction enzyme-related protein	No significant hit		Deltaproteobacteria; *Anaeromyxobacter dehalogenans* 2CP-C	2.0 e^{-13}	51.32	Q2IGC5_ANADE
255	Bacteriophage-related protein of unknown function	*P. fluorescens* SBW25	1.0 e^{-20}	dsDNA virus; *Pseudomonas* phage F116	2.0 e^{-29}	37.87	Q5QF30_9CAUD
289	Plasmid conjugal transfer inhibition protein, Tir-like	No significant hit		Gammaproteobacteria; *Erwinia amylovora* (pEL60)	1.0 e^{-12}	32.81	Q6TFZ5_ERWAM
301	Ankyrin repeat-containing protein	No significant hit		Eukaryota; *Mus musculus*	2.0 e^{-12}	34.52	Q8C8R3_MOUSE
307	Acetyltransferase GNAT family protein	Top hit		Gammaproteobacteria; *P. syringae* 1448A	1.0 e^{-51}	45.33	Q48GQ0_PSE14
311	Stringent starvation protein, SspA-like homologue	*P. syringae* 1448A	1.0 e^{-42}	Gammaproteobacteria; *Halorhodospira halophila* SL1	5.0 e^{-21}	32.64	Q2CS52_ECTHA
318	Plant-inducible oligoribonuclease, Orn	No significant hit		Eukaryota; *Tetrahymena thermophila* SB210	2.0 e^{-29}	42.94	Q22ZB0_TETTH
344	Transcriptional regulator, AlgZ-like homologue	*P. fluorescens* Pf-5	1.0 e^{-18}	Gammaproteobacteria; *Azotobacter vinelandii* AvOP	5.0 e^{-19}	51.55	Q4J155_AZOVI
350	RNA polymerase sigma-32 factor, RpoH-like homologue	Top hit		Gammaproteobacteria; *P. aeruginosa*	2.0 e^{-84}	62.45	RP32_PSEAE
361	Plant-inducible DNA helicase, HelB	Top hit		Gammaproteobacteria; *P. resinovorans* (pCAR1)	0	54.86	Q8GHN5_PSERE
364	Cold shock DNA-binding domain protein, Csp-like homologue	Top hit		Gammaproteobacteria; *P. putida* KT2440	7.0 e^{-41}	39.18	Q88Q61_PSEPK
367	DnaJ family protein	*P. fluorescens* SBW25	8.0 e^{-08}	Gammaproteobacteria; *Nitrosococcus oceani* ATCC 19707	5.0 e^{-15}	54.29	Q3JC07_NITOC
371	DNA helicase	*P. aeruginosa*	7.0 e^{-11}	Betaproteobacteria; *Burkholderia vietnamiensis* G4	2.0 e^{-56}	32.85	Q4BMY9_BURVI

Table continued overleaf

377	Transcriptional regulator, AlgZ-like homologue	Top hit	5.0 e^{-04}	Gammaproteobacteria; *P. aeruginosa*	5.0 e^{-04}	46	Q9RPY7_PSEAE
383	Plasmid IncA/C–Inc P3 replication protein, RepA	No significant hit		Gammaproteobacteria; *Escherichia coli* (pRA1)	1.0 e^{-61}	44.56	Q08896_ECOLI
401	DNA helicase	Top hit		Gammaproteobacteria; *P. syringae* B728a	7.0 e^{-130}	48.28	Q4ZWH0_PSEU2
403	DNA restriction methylase	Top hit		Gammaproteobacteria; *Pseudomonas* sp. ND6 (pND6-1)	1.0 e^{-72}	35.87	Q6XUK5_9PSED
407	DNA ligase, bacteriophage-like homologue	No significant hit		dsDNA viruses; Bacteriophage KVP40	6.0 e^{-28}	27.52	Q6WI94_BPKV4
413	Single-strand binding protein, Ssb-like homologue	Top hit		Gammaproteobacteria; *P. syringae* DC300	3.0 e^{-29}	38.74	SSB_PSESM
421	Response regulator receiver domain protein	Top hit		Gammaproteobacteria; *P. aeruginosa*	1.0 e^{-34}	53.23	PILG_PSEAE
426	Tn5042-like transposase, TnpA	Top hit		Gammaproteobacteria; *P. fluorescens*	2.0 e^{-61}	99.15	Q70MR7_PSEFL
427	Tn5042-like transposase, TnpB	Top hit		Gammaproteobacteria; *P. fluorescens*	1.0 e^{-61}	100	Q70MR8_PSEFL
428	Tn5042-like transposase, TnpC	Top hit		Gammaproteobacteria; *P. fluorescens*	0	99.4	Q70MR9_PSEFL
430	Tn5042-like organomercurial lyase, MerB	Top hit		Gammaproteobacteria; *P. fluorescens*	2.0 e^{-115}	99.06	Q70MS1_PSEFL
431	Tn5042-like mercuric ion reductase, MerA	Top hit		Gammaproteobacteria: *P. fluorescens*	0	96.25	Q70MS2_PSEFL
432	Tn5042-like inner membrane mercury ion uptake protein, MerC	Top hit		Gammaproteobacteria; *P. fluorescens*	1.0 e^{-55}	93.75	Q53IQ9_PSEFL
433	Tn5042-like periplasmic mercuric ion binding protein, MerP	Top hit		Gammaproteobacteria; *P. fluorescens*	9.0 e^{-44}	100	Q53IR0_PSEFL
434	Tn5042-like mercuric ion transport protein, MerT	Top hit		Gammaproteobacteria; *P. fluorescens*	3.0 e^{-43}	99.14	Q53IQ8_PSEFL
435	Tn5042-like Mer operon activator/repressor, MerR	Top hit		Gammaproteobacteria; *P. fluorescens*	1.0 e^{-76}	99.3	Q70MS3_PSEFL

Table continued overleaf

438	Recombination-associated protein, RdcG-like homologue	Top hit		Gammaproteobacteria; *P. fluorescens* Pf-5	4.0 e^{-93}	61.76	Q4K8D9_PSEF5
443	Carbon storage translational RsmA/CsrA family regulator, RsmA-like homologue	Top hit		Gammaproteobacteria; *P. fluorescens* Pf-5	4.0 e^{-14}	73.08	Q4KEY0_PSEF5
445	Plasmid IncA/C–Inc P3 replication protein, RepA	No significant hit		Gammaproteobacteria; *Buchnera aphidicola* (pBPs2)	1.0 e^{-33}	29.61	Q9ZER8_9ENTR
455	Plasmid partitioning protein, ParB	*P. syringae* 1448A	1.0 e^{-18}	Betaproteobacteria: *Ralstonia metallidurans* CH34	1.0 e^{-61}	30.88	Q5NUW9_RALME
461	Exodeoxyribonuclease I	*P. putida* KT2440	1.0 e^{-10}	Gammaproteobacteria; *Methylococcus capsulatus*	4.0 e^{-15}	24.77	Q605A2_METCA
464	DNA polymerase III subunit	*P. syringae* 1448A	4.0 e^{-31}	Gammaproteobacteria; *Vibrio fischeri* ATCC 70601	1.0 e^{-31}	28.53	Q5E8Z1_VIBF1
465	RNA polymerase sigma factor, RpoD-like homologue	*P. putida* F1	5.0 e^{-17}	Betaproteobacteria; *Ralstonia solanacearum*	2.0 e^{-20}	23.95	Q8XXA1_RLSO
470	Chemotaxis protein CheB like Methylesterase	*P. aeruginosa*	3.0 e^{-13}	Gammaproteobacteria; *Nitrosococcus oceani* ATCC 19707	2.0 e^{-19}	31.42	Q3JET8_NITOC
471	Two-component regulator histidine kinase fused response regulator protein	*P. aeruginosa*	9.0 e^{-73}	Gammaproteobacteria; *Alkalilimnicola ehrlichei* MLHE-1	5.0 e^{-78}	35.27	Q34VJ7_9GAMM
472	Chemotaxis protein, CheR Methyltransferase	Top hit		Gammaproteobacteria; *P. aeruginosa*	2.0 e^{-24}	31.58	Q51346_PSEAE
473	Methyl-accepting chemotaxis Sensory transducer protein (MCP)	*P. fluorescens* SBW25	2.0 e^{-22}	Gammaproteobacteria; *Alkalilimnicola ehrlichei* MLHE-1	8.0 e^{-25}	24.06	Q34VJ8_9GAMM
474	Chemotaxis signal transduction protein CheW	No significant hit		Gammaproteobacteria; *Xylella fastidiosa*	2.0 e^{-6}	27.66	Q9PC31_XYLFA
475	Two-component response regulator CheY like reciever protein	*P. aeruginosa*	3.0 e^{-17}	Gammaproteobacteria; *Psychrobacter arcticum*	1.0 e^{-19}	42.98	Q4FQP2_PSYAR
477	Plasmid partitioning protein, ParB	*P. putida* F1	9.0 e^{-29}	Gammaproteobacteria; *Legionella pneumophila* Lens	2.0 e^{-36}	35.23	Q5WTL1_LEGPL

Table 2.3 of the pQBR103 CDS which have been functionally inferred and their blastp (Altschul et al., 2001) (version = 2.2.12 matrix = Blosum 62, word length= 3, gap penalties existence 11, extension, 1) homology to a custom database, comprising the entire Uniprot database (www.ebi.uniprot.org/index.shtml) (date 3/6/2006) and the *P. fluorescens* SBW25 proteome (Unpublished). Closest hit for each CDS encoded protein to the database is given Where this is not to a Pseudomonad plasmid or chromosomal sequence, the closest hit to a Pseudomonas sequence is also given. No significant homology indicates the CDS has no homology to a Pseudomonad sequence. Low homology, refers to weak levels of homology but over reasonable length of the sequence and annotation is supported by its genomic context.

* Annotation given is based on inspection of homology over the entire length of the CDS, using Blastp and Fasta against Uniprot and Pfam (see material and methods).

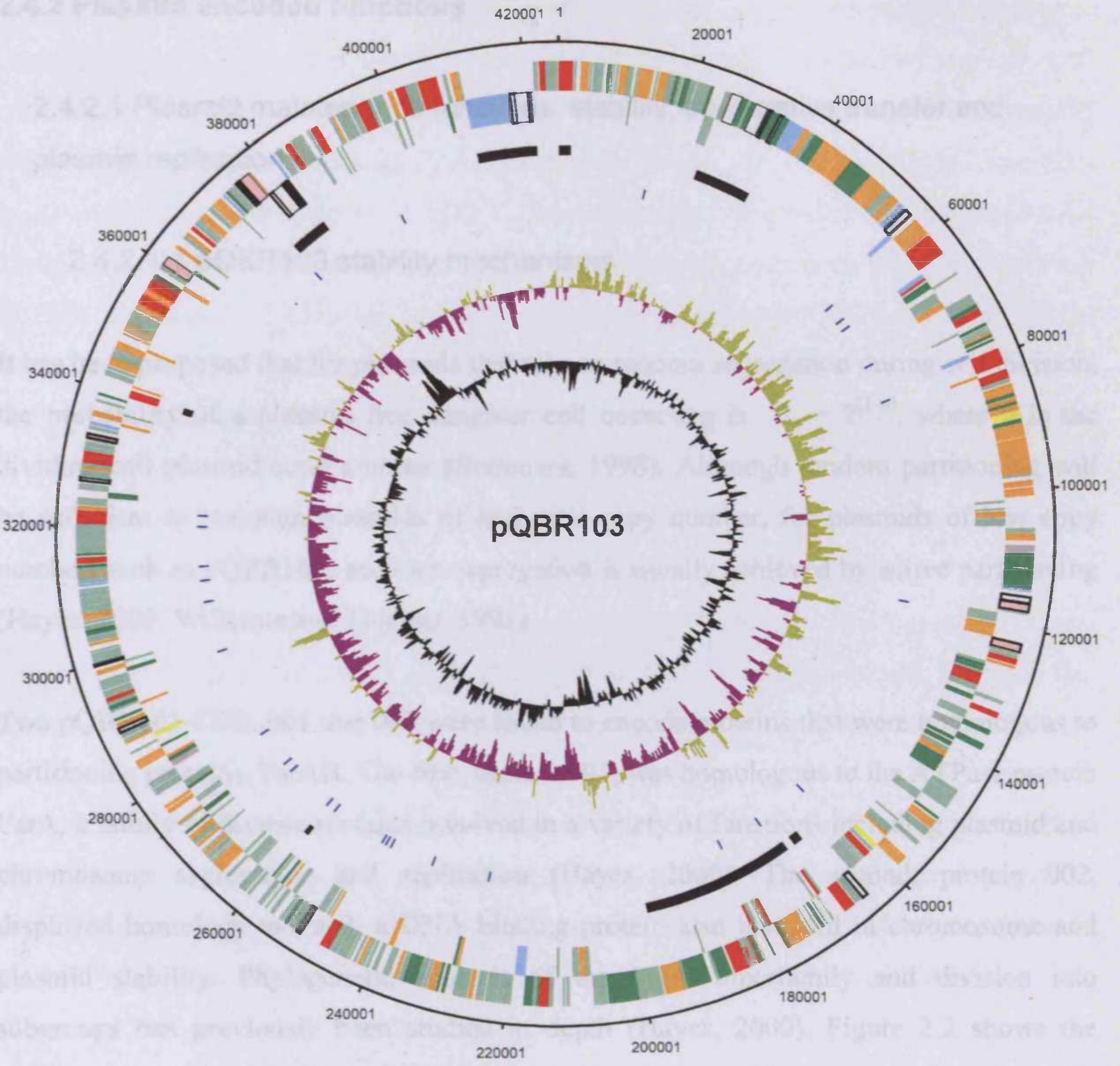


Figure 2.1. General features of the pQBR103 genome. Outermost circle, **Circle 1** (Black) represents the pQBR103 sequence, numbers refer to sequence position (bp). Note position 0 on diagram is arbitrary and not the annotated origin of replication, which could not be ascribed. **Circle 2 and 3** (multicoloured) CDS sequences annotated to the forward and reverse strands respectively, colours refer to CDS categorisation (Functional CDSs: Red, DNA associated; Yellow, metabolism; Pink, phage or transposon; White, environmental survival / transmission; Blue, regulatory, Grey, domain match only. Conserved Hypotheticals CDSs, Orange. Orphans, Light green. Transmembrane predicted CDSs, Dark green, note these CDSs are a mixture of Orphan, Conserved hypotheticals and functional characterised CDS). **Circle 4** (Black blocks) regions on pQBR103 of maintenance function or other interest (clockwise from pQBR103 sequence position 1) 1, plasmid partitioning region; 2, region of transmebrane/secretion homologues; 3, UV resistance genes; 4, region covering Type IV transfer like homologues; 5, region of homology to *oriV* of pQBR11; 6, replication initiation protein; 7, Tn*5042* like transposon and mercury resistance operon; 8, region of chemotaxis like homologues. **Circle 5** (Blue blocks) Regions of IVET insertions where potential plant-induced transcription may occur. **Circle 6** (gold and purple) G+C skew, **Circle 7** (black) GC deviation

2.4.2 Plasmid encoded functions

2.4.2.1 Plasmid maintenance functions: stability, conjugative transfer and plasmid replication

2.4.2.1.1 pQBR103 stability mechanisms

It has been proposed that for plasmids that rely on random segregation during cell division, the probability of a plasmid free daughter cell occurring is $P_o = 2^{[1-n]}$, where n is the dividing cell plasmid copy number (Summers, 1998). Although random partitioning will be sufficient to maintain plasmids of high cell copy number, for plasmids of low copy number, such as pQBR103, accurate segregation is usually achieved by active partitioning (Hayes, 2000; Williams and Thomas, 1992).

Two pQBR103 CDS, 001 and 002 were found to encode proteins that were homologous to partitioning proteins, ParAB. The first, protein 001, was homologous to the ATPase protein ParA, a family of diverse proteins involved in a variety of functions including plasmid and chromosome segregation and replication (Hayes, 2000). The second, protein 002, displayed homology to ParB, a DNA binding protein also involved in chromosome and plasmid stability. Phylogenetic analysis of the ParA superfamily and division into subgroups has previously been studied in depth (Hayes, 2000). Figure 2.2 shows the phylogenetic analysis of the pQBR103 ParA homologue with representatives from the ParA superfamily subgroups, many of which have been functionally determined to be involved in active partitioning. Notably, pQBR103 was grouped with the IncC subgroup, although only at the basal level. This is of interest as the ParA homologues of IncC plasmids including R751 and RK2 are not solely confined to partitioning, also being involved in the regulation of vegetative replication and conjugation (Bignell and Thomas, 2001). The operon structure of R751 and RK2, are not similar to that found on pQBR103. The gene organisation and orientation of pQBR103 ParAB homologues closely resembled that of the P1 and F plasmid partition regions (Abeles et al., 1985; Mori et al., 1986). Therefore, it might be expected that the proposed partition mechanism would more likely resemble these systems.

The mechanism for active partitioning is still not completely understood although a popular model is that ParB binds to a *cis* acting centromere-like region termed *parS*, resulting in plasmid pairing. This complex then separates towards the cell poles possibly involving host bacterial machinery (Williams and Thomas, 1992). In pQBR103 iterations of a degenerate 6 bp repeat (TGCTTT) were identified downstream from the ParAB homologues and may constitute a *parS* region. In the partitioning mechanism the exact role of ParA is still not fully known, however, disruption of this gene in P1 has resulted in the loss of plasmid stability (Davis et al., 1992). It has also been shown that ParA's ATPase activity is stimulated by the presence of the ParB protein and this activity may be involved in the separation of the paired plasmid complex prior to moving towards the cell poles (Williams and Thomas., 1992). In addition, some ParA's act as site specific binding proteins that bind upstream of ParA and B and, therefore, could possibly be involved in regulation of the operon. Although the aforementioned ParA and ParB homologues, and putative *parS* region, identified in pQBR103 would fit within this model, whether this region is responsible for partitioning of the plasmid is yet to be experimentally determined.

Of note from annotation of pQBR103, was the identification of three other ParB homologues (encoded by CDSs 113, 455 and 477), that were not genetic homologues of the aforementioned ParB (encoded by CDS 002). However, none of these were found in close proximity to a ParA homologue to suggest a candidate active partitioning system. They may therefore be involved in a regulatory role.

In addition to a putative active partitioning system, pQBR103 was predicted to encode proteins that may be involved in the resolution of plasmid dimers. Dimers are formed from homologous recombination events between sister plasmids within a cell. Dimer resolution systems are a particular feature of high copy number plasmids, which have no active partitioning systems and therefore rely on segregation by a random process at cell division. Dimer resolution in such plasmids is important because dimers contain duplicate replication inhibition genes, effectively meaning inhibition is two fold higher compared to monomers. This means the number of dimer replicons in a cell is half that compared to monomer replicons. Therefore, the accumulation of dimers in a cell increases the frequency of plasmid free daughters at cell separation (Summers, 1993). However, even low copy number plasmids such as pQBR103, and chromosomes, need to be resolved to ensure accurate separation at division and avoid the burden of dimers. pQBR103 was found to encode two putative site specific recombinases of the λ intergrase family (CDS 105 and 110). These proteins showed homology to XerC and XerD proteins. These XerCD proteins have been shown in plasmid ColE1 to form a heterodimer at specific *cer* sites, which with the involvement of chromosomally encoded proteins, ArgA and PepA, mediate site specific recombination and consequently dimer resolution to monomers (Summers, 1998). Whether a similar system is active in pQBR103 can only be inferred.

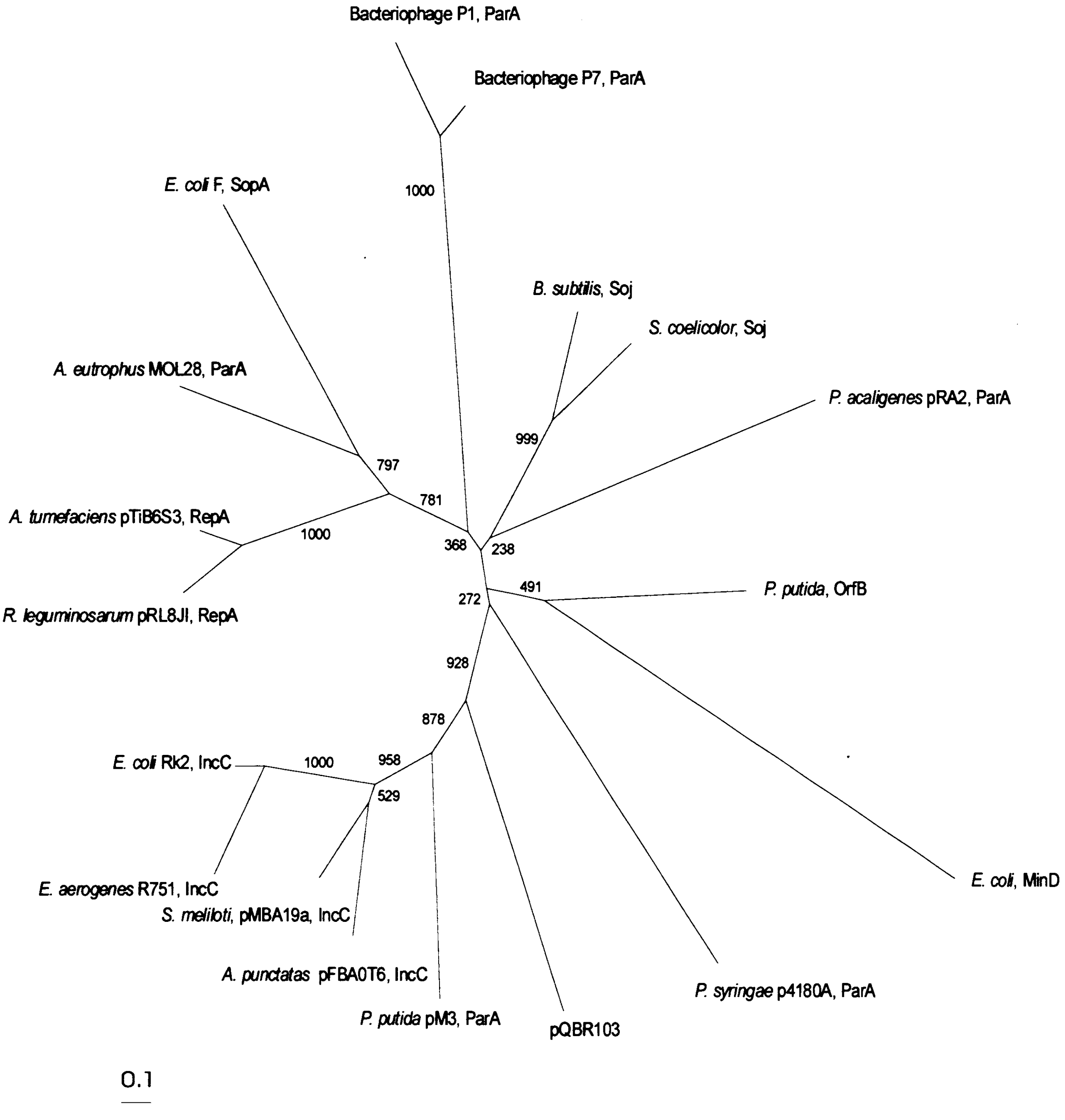


Figure 2.2. Phylogenetic tree of the ParA homologue encoded by pQBR103 CDS 001, in relation to closest Blastp homologues as well as to a diverse range of functionally characterised proteins of the ParA superfamily. Alignments were made in ClustalX and tree constructed using a Neighbor-Joining method. Each tree was bootstrapped (1000 replicates).

2.4.2.1.2 Conjugative transfer

As introduced in chapter 1, probably the most important driver of plasmid transfer and indeed HGT in natural environments, particularly the phytosphere (Davison, 1999; van Elsas et al., 2003), is conjugation. The pQBR plasmid collection as a whole was shown to be conjugative, confirmed by laboratory tests. Notably, it has been shown that the transfer proficiency of these plasmids *in situ* is correlated with plant maturation, stimulated by unknown phytosphere factors (Bailey et al., 2001).

Conjugation involves two processes; firstly, mating pair formation for cell to cell contact, mediated by a type IV system, and secondly DNA processing and transfer. These systems are brought together by a coupling protein. Diversity of these systems has been observed in different Gram negative plasmids, in terms of the genes they encode as well as their genomic organisation i.e. sometimes they are clustered in a single region (e.g. the IncF1 F factor type) (Frost et al., 1994) or sometimes they are functionally clustered in two regions (e.g. the IncPβ R751) (Thorsted et al., 1998). Nevertheless, all described conjugative plasmids tend to display shared homology for a substantial number of proteins and can be ascribed to this paradigm/model of transfer (Scroeder and Lanka, 2005).

In initial screens of the pQBR103 genome (see materials and methods), two regions of contiguous CDS (that are likely to be co-transcribed i.e. are putative operons), were identified that are predicted to encode proteins homologous to known, or putatively ascribed Type III/IV or other conjugation related proteins. The first region spans CDS 27-36 (sequence position: 24398-34265) and the second, CDS 470 –475 (seq position: 410934..421365). In addition, a third region spanning 27 kb, CDS 160 –191 (seq position 170292-197117), that are unlikely to be co-transcribed, was identified that contains a number of potential Dtr or CP/Mpf like homologues (Table 2.5).

On closer inspection of the second region (CDS 470-475) it is probable this does not participate in plasmid conjugation, but instead, likely has role in bacterial chemotaxis (see section 4.2.2.2.2). Of the other two remaining candidate transfer regions, region 3 contained the largest number of homologues that could be putatively ascribed a function in CP/Mpf by a Type IV mechanism, or Dtr. In Table 2.5, pQBR103 CDS that are potential functional homologues to the characterised IncF1 F plasmid conjugation system or the pTiC58 T-DNA Type IV system is shown. It is proposed that this represented the most likely candidate region for the CP/Mpf functioning. However, whether this also includes the Dtr system is less clear. A putative DNA primase (CDS 209) and transfer inhibition protein (CDS 289) which might be expected to participate in the conjugation process and be located near to the origin of transfer was found to lie outside this region.

Conjugation processes in Gram-negative bacteria associated plasmids involves ~20-35 genes. For a functional type IV secretion system alone this requires 8-14 proteins, based on the analysis of functionally characterised systems (IncF 1 F plasmid Tra; pTiC58 VirB/D4; and IncPβ R751 Trb). Based on the analysis of the aforementioned 27 kb region (candidate region 3) there were insufficient genes to ascribe this region as either responsible for conjugation, or containing a functional Type IV system. Further functional investigation of this region, and region 1, by gene disruption would be required to implicate or disprove the involvement of these regions in conjugation. What is clear from the annotation of pQBR103 is that the mechanism of transfer remains largely elusive, and at best is likely to have a system that is highly divergent from other characterised plasmid conjugation systems.

			Closest hit		
pQBR103 CDS	Annotation	VirB/D and IncF1Tra functional homologue[§]	Id/similarity %	E-value	Accession
160	Conjugal transfer TrbN-like homologue	VirB1/ F Orf169[†]	32/54	7e-3	YP_190354
175	Transcriptional regulator, TraR/DskA-like homologue	-	46/63	2e-21	ZP_01639571
181	Exodeoxyribonuclease/helicase RecD/TraA-like homologue	-	21/37	5e-16	ZP_01144634
182	Conjugal transfer TraG/D family protein, Coupling protein homologue	VirD4TraG/ TraD/[†]	25/43	2e-33	ZP_00242688
188	Conjugal transfer TraB-family protein	VirB10/TraB[†]	26/40	1e-05	AAW83066
189	Conjugal transfer TraV like homologue	VirB7/TraV*	33/43	1e-4	YP_096101
191	Conjugal transfer TraC-like homologue	VirB4/TraC*	23/42	2e-33	NP_640173

Table 2.5. pQBR103 CDS contained in the 27 kb region (pQBR103 genome position 170292-197117) that had homology to proteins with functionally determined or predicted roles in either conjugation and/or type IV secretion assembly. Closest hit refers to the closest homologue in the NCBI nr protein database described with a putative conjugative/type IV function. (using the blastp program (version = 2.2.12 matrix = Blosum 62, word length = 3, gap penalties existence 11, extension, 1) (date of search = 20/11/2006) [§]Possible functional homology to the paradigm Type IV systems, the IncF1 F factor and pTiC58 systems, are given (GenBank accession U01159, AE007871 respectively). *denotes direct genetic homology to one or both systems (Blastp expected value of < 0.01), [†] denotes indirect genetic homology to one or both systems via intermediates.

2.4.2.1.3 Plasmid replication

Autonomous replication from the host chromosome is a defining characteristic of plasmids, although it does rely to differing degrees on host machinery. The process of plasmid replication, however, must be controlled and be responsive to host cell division. If replication rate is too high this will cause an unnecessary burden, if too low (less than once per host generation), plasmid free daughter cells will accumulate upon cell division.

In a previous study a 300 bp minimal origin of replication was identified in pQBR11 (Viegas et al., 1997), a group I plasmid shown to be closely related to pQBR103 (based on RFLP analysis and reciprocal hybridisations) (Lilley et al., 1996). Plasmid pQBR103 was shown to share this region with pQBR11 (pQBR103 sequence position: 259,339–259,639), with the exception of a single nucleotide insertion. It is not believed that this region is a functional minimal replicon in pQBR103, due to lack of associated replication features in close proximity, suggesting this is an additional *oriV*. In total two homologues of replication initiator proteins were found to be encoded on pQBR103 (CDS 0383 and 0445), with reciprocal homology to each other. Both homologues showed homology to the RepA IncA/C-IncP3 of the 129 kb RA1 plasmid (Llanes et al., 1996). Phylogenetic analysis of both pQBR103 RepA homologues is shown in Figure 2.3. Only CDS 0383 had associated repeats (thirty-two copies of a 22 bp repeat (GTTGTAGGTTTG(A/G)TG(G/C)GCCCTA) and two DnaA-boxes to suggest it may represent a functional replicon similar to that described for RA1 (Llanes et al., 1996). Therefore, it is proposed (although not experimentally determined) that pQBR103 replicates bidirectionally to form a so called theta intermediate. The mechanism and control of this replication maybe similar to the model described for iteron-containing replicons (Espinosa et al., 2000).

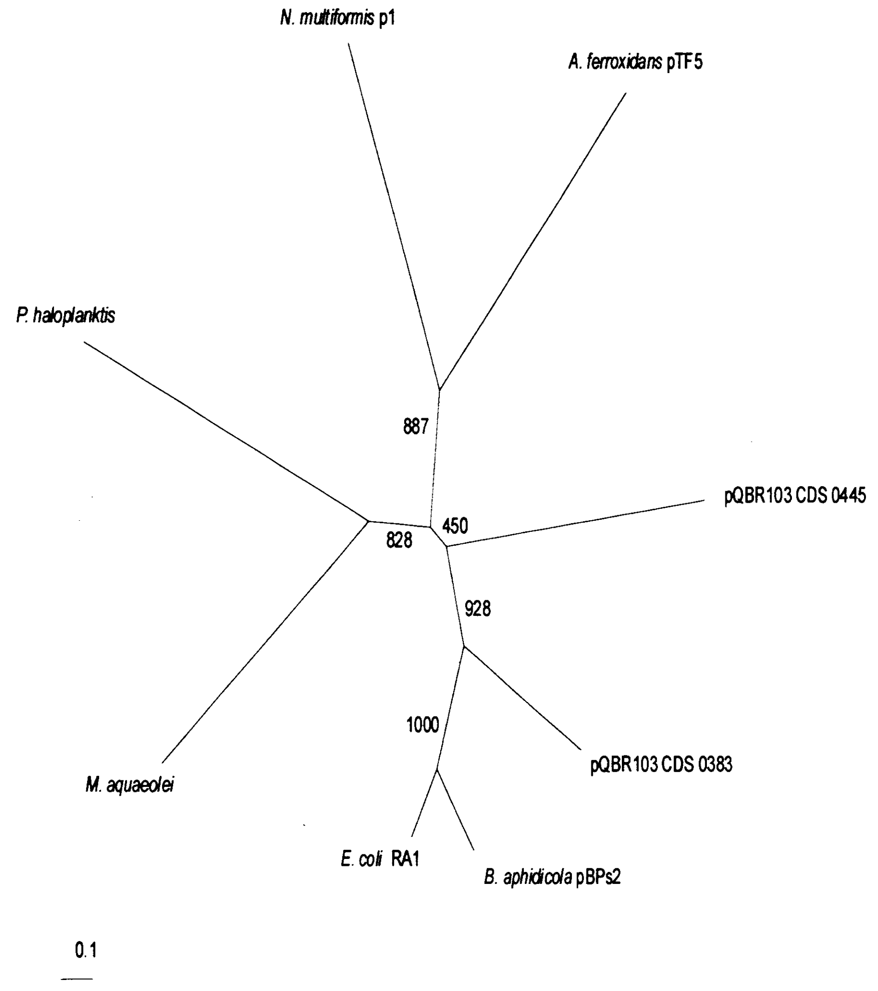


Figure 2.3. Phylogenetic tree of the RepA homologues encoded by pQBR103 CDS 383 and 445, in relation to all Blastp homologues. Alignments were made in ClustalX and tree constructed using a Neighbor-Joining method. Each tree was bootstrapped (1000 replicates).

2.4.2.2 Functions of putative ecological relevance

2.4.2.2.1 Mercury resistance

Mercury compounds are highly toxic to all living organisms, yet mercury is near ubiquitous in the natural environment on a world wide scale (Dahlburg et al., 1997). Analysis of sediment ice cores (Schuster et al., 2002) has shown an increase in the mercury input to the environment due to industrial processes. However, it is geological processes that are responsible for a large amount of environmental mercury, and it has been suggested that huge quantities were released due to ancient volcanic activity (Osborn et al., 1997). Bacteria have evolved a number of different mercury resistance mechanisms many of which are borne on MGEs (Osborn et al., 1997).

Due to mercuric resistance being a common selectable trait of MGE in natural systems, inorganic mercuric resistance was used as the selection criterion for the study of plasmid abundance at the Wytham field site. Subsequent analysis of this plasmid collection has identified two distinct *mer* types. The first common to pQBR groups I, III and IV, and the second common to groups II and V (A. Lilley, personal communication). Sequencing pQBR103 (group I) has revealed the probable mechanism for this resistance and acquisition of this trait by transposition. Upstream of the *mer* operon there three transposase genes were identified (*tnpA, tnpB, tnpC*). Flanking both the *mer* operon and transposases were imperfect inverted repeats, indicating a type II composite transposon structure (Figure 2.4). The whole transposon, from inverted repeat to inverted repeat had the same gene order to that of Tn*5042*, a transposon sequenced from a *P. fluorescens* strain isolated from permafrost samples (15-40 thousand years old) in Russia (Mindlin et al., 2005). This suggests this mer type is ancient and was widely distributed prior to the impact of industrial pollution (Mindlin et al., 2005). Comparison of the pQBR103 transposon-like region and Tn*5042* revealed they are near sequence identical and Tn*5042* was pQBR103's closest database hit (Blastn id 6941/6991 (99%)). In phylogenetic analysis, the MerA of Tn*5042* was found to be the closest neighbour to the MerA homologue of pQBR103 (encoded by CDS 0431). This data is not shown, since the MerA of Tn*5042* in relation to other *mer* types has been analysed previously (Vetriani et al., 2004).

The *mer* operon of pQBR103 was found to contain a common set of *mer* genes (*merRTPCA*). In addition it contained a *merB* homologue, which encodes an organomercuric lyase, enabling broad spectrum resistance i.e. to both organic and inorganic mercury. The mechanism of mercuric resistance has been reviewed previously (Osborn et al., 1997; Barkay et al., 2003) and is summarised in Figure 2.5. Testing pQBR103 and other pQBR plasmid groups (representing group I, III and IV) believed to share a common mer type, it was demonstrated that all these plasmids conferred broad spectrum mercuric resistance (Table 2.6).

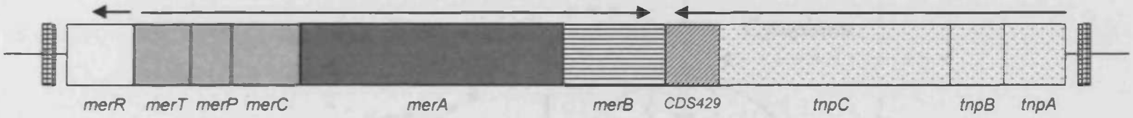


Figure 2.4. Type II composite mercury transposon structure (Tn*5042* like) on pQBR0103. Transposase genes are indicted by dotted boxes. Light grey boxes indicate mer genes with a transport function; dark grey, reductase function; open box, regulatory function; and horizontally lined box, lyase function (mechanism of mercury resistance is described in figure 2.5). The diagonally lined box represents a conserved hypothetical pseudogene. Checked boxes indicate imperfect inverted repeats. Direction of transcription is indicated by arrows.

	Phenylmercuric acetate (μg ml^{-1})				**Mercury chloride** (μg ml^{-1})		
	10	5	2	0.2	27.2	5.44	2.72
Strain or plasmid (group)							
P. putida UWC1	none	1000x inhibit	1000x inhibit	full	none	none	none
pQBR4 (I)	10x inhibit	full	full	full	full	full	full
pQBR41 (I)	full	full	full	full	full	full	full
pQBR42 (I)	full	full	full	full	full	full	full
pQBR44 (I)	full	full	full	full	full	full	full
pQBR47 (I)	full	full	full	full	full	full	full
pQBR55 (III)	full	full	full	full	full	full	full
pQBR29 (III)	full	full	full	full	full	full	full
pQBR57 (IV)	full	full	full	full	full	full	full

Table 2.6. Mercuric resistance levels of different pQBR plasmids which share a common *mer* type. Full refers to complete cell growth at the maximum cell dilution (10^{-7}). All plasmids showed synonymous levels of resistance to both PMA and $HgCl_2$ other than pQBR4.

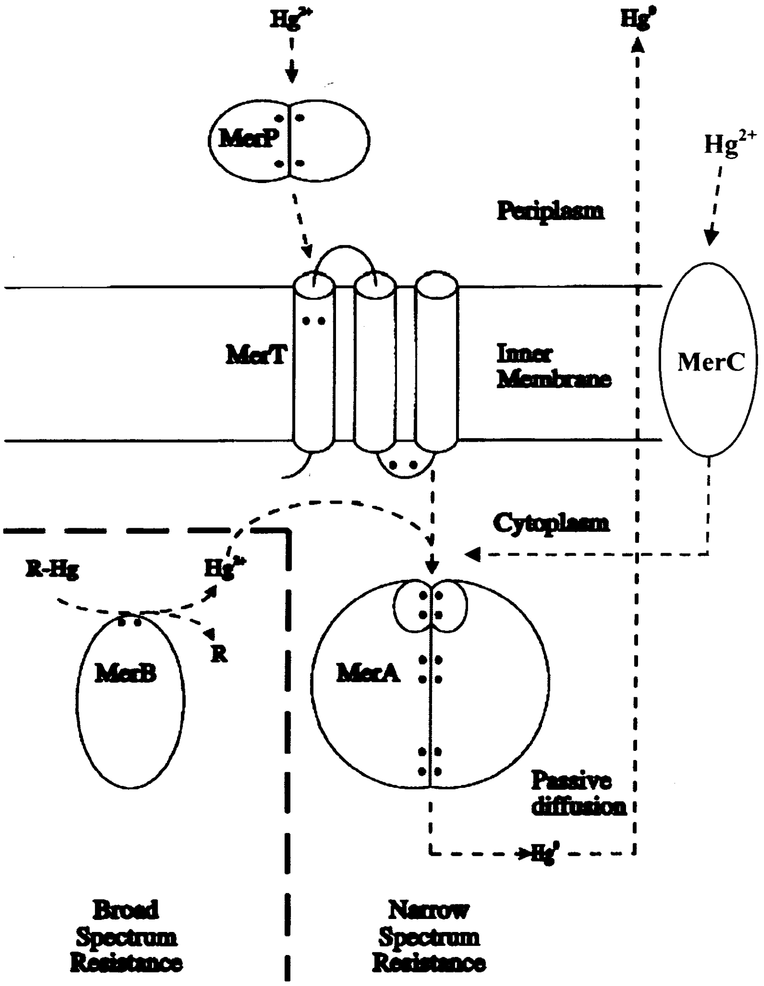


Figure 2.5 The proposed mechanism of narrow and broad spectrum mercury resistance (image modified from Osborn et al., 1997 to show involvement of MerC protein). The mechanism of resistance is as follows: MerP is a small periplasmic protein that sequesters Hg^{2+} ions. These mercuric ions are then transferred to the transmembrane protein MerT which transfers the ions to the cytosol. Alternatively, ions can become available by organomercuric lyase (MerB) that transfers a proton to the C-Hg bond yielding the Hg^{2+}. Once in the cytosol ions are reduced by the activity of the mercury reductase, MerA. Volatile lipid soluble Hg^0 is then free to diffuse from the cell. This whole process is under the regulation of MerR. In addition to the basic model described pQBR103 encodes a further protein, MerC. MerC homologues are not always found in bacterial mercury resistance operons and the function of this protein is not fully understood. However, it is thought MerC is a transmebrane protein that can transfer Hg^{2+} into the cell cytoplasm without the involvement of MerP (Barkay et al., 2003).

2.4.2.2.2 Other putative ecological functions

Another pQBR103 encoded trait includes resistance to ultra violet (UV) radiation. In natural environments bacteria are faced with a myriad of environmental stresses including DNA damage induced by solar UV radiation. This is particularly true for Bacteria of the phyllosphere. Here Bacteria reside on the principle photosynthetic organs which are arranged to maximise exposure to sunlight, and hence are exposed to high levels of UV radiation (Sundin and Murillo, 1999). In such an environment UV resistance is of considerable benefit to host. Numerous studies have investigated the distribution of UV resistance determinants in natural environments and found that they are frequently partitioned in the HGP. For example, UV resistance was found to be a common trait of the large pseudomonas plasmids isolated from the River Taff epilithon (Cardiff South Wales) (Fry and Day, 1990). Similarly, it has been demonstrated that UV resistance is common to the pPT23A-like plasmids indigenous to the phytosphere *P. syringae* pathovars (Sundin and Murillo, 1999).

In previous analyses of pQBR103 it was demonstrated that the plasmid confers UV resistance to *P. fluoreccens* SBW25. This resistance was found to be a common phenotype to all pQBR group I plasmids tested (Zhang et al., 2004b). Analysis of the pQBR103 genome revealed two genes *rulA* and *rulB* (CDS 157 and 158) which were homologous to genes which have been functionally determined to be involved in resistance to UV (Sundin and Murillo, 1999), and are thus predicted to account for the observed UV resistance.

In addition to UV resistance, pQBR103 was shown to encode a number of putative proteins homologous to regulatory proteins, and as such may control plasmid/host responses to the environment. There was a homologues of a carbon storage regulator of the CsrA/RsmA family (CDS 443), a protein family that is implicated in the repression of stationary/secondary metabolite genes (Gutierrez et al., 2005). In addition, two homologues (CDS 056, 073) to catabolite Crp/Vfr global regulators, and two putative RNA polymerase sigma factors (CDS 350, 465). Two putative proteins that contain conserved GGDEF domains where also found (CDS 53 and 213). These domains have found to be present in hundreds of bacterial species and are implicated in the synthesis of cyclic diguanylate (c-di-GMP), an important second messenger in bacterial cells (Romling et al., 2005)

Another putative role that was ascribed to pQBR103 is chemotaxis. Chemotaxis is the movement of bacteria towards favourable environments, whether that is towards chemoattractants (positive chemotaxis) or away from chemorepellents (negative chemotaxis) (Pandey and Jain, 2002), usually achieved by regulation of flagella rotation. For free living bacteria chemotaxis provides a mechanism to locate carbon sources (Grimm and Harwood., 1999). For plant associated bacteria, chemotaxis is recognised as an important trait for sensing plant exudates such as simple sugars, amino acids and plant flavonoids, leading to the colonisation and survival in the plant rhizosphere (de Weert et al., 2002; Pandya et al., 1999), and also may be relevant to survival and colonization of the phyllosphere (Beattie and Lindlow, 1995). Although chemotaxis determinants are commonly found on the chromosomes of environmental bacteria, plasmids are also frequently associated with roles in chemotactic responses (Grimm and Harwood, 1999; Pandya et al., 1999).

Annotation of pQBR103 revealed six CDS that may be co-transcribed (CDS 470-75). The six proteins putatively encoded were homologous to proteins putatively or functionally ascribed to be involved in chemotaxis. Sufficient homologues were found in pQBR103 to suggest a complete two component signal transduction system, likely to be involved in chemotaxis. Such systems are well conserved in Bacteria and the Archaea; the mechanisms of these systems are described in the reviews by Szurmant and Ordal, (2004) and Lux and Shi (2004). Briefly, the chemotactic signal transduction pathways can be divided into three basic elements; Firstly, signal reception by proteins usually on the membrane; secondly, signal transduction from receptors to the flagella motor complex; thirdly, signal adaptation in order to desensitise the initial input (Lux and Shi, 2004).

Based on homology searches the following functions for CDS 470-475 are proposed. Firstly CDS 473 putatively encodes a methyl accepting protein (MCP) and is the likely receptor of the system (first element of the system). The signal in turn is transduced to the receptor bound protein encoded by CDS 474 (CheW homologue) which couples the protein derived from CDS 471 (Containing a conserved CheA domain), a putative histidine kinase, to the receptor. The putative histidine kinase in turns transduces the signal to the protein of CDS 470 (CheB like homologue), a putative methyl esterase, and that of CDS 475 (a CheY homologue), a putative response regulator. The CheY homologue is predicted to directly interact with the flagella motor complex (completing the second element of the system). The third element of the system, adaptation to desensitise the original input, is likely achieved by the proteins of CDS 470 (the aforementioned CheB homologue) and CDS 472 (CheR like homologue), a putative methyl transferase. These proteins are predicted to be antagonistic to one another with the protein of CDS 470 (CheB like), once activated by the putative histidine kinase (CDS 471), demethylating the chemoreceptor to afford it less sensitivity, and the protein of CDS 472 (CheR like) methylating it to increase sensitivity (Szurmant and Ordal, 2004).

Chemotaxis activity has never been demonstrated or indeed investigated for pQBR103, to the author's knowledge. Based on the identification of this homologous system, assays could be devised to investigate the response conferred on potential natural hosts by pQBR103, to plant exudate components. If this is a functioning chemotaxis signal transduction system this offers further understanding of the ecological role of this plasmid.

2.4.3 Mapped *in vivo* expressed sequences

IVET technology has been used previously to identify regions of pQBR103 that contain promoters induced in the immature sugar beet phytosphere; subsequent analysis of some of the IVET fusions and the genes contained within them have been reported (Zhang et al., 2004a;b). In this study these published IVETs and those publicly accessible were mapped on to the pQBR103 plasmid genome (28 in total). Their distribution is shown in Figure 2.1. From these IVETs it is possible to give an estimation of the number and composition of genes that may be potentially controlled by an environmentl promoter. In total, 11 regions were identified that could contain environmentally transcribed genes (Table 2.7). These regions suggest that up to 74 pQBR103 genes may be transcribed *in planta* on sugar beet seedlings. This represents 15.5% of the annotated genetic capacity of the plasmid. Of these 74 genes only 12 (16%) were inferred a function and 10 (14%) were conserved hypotheticals. However, the vast majority (70%) were orphan sequences, and as such little can be inferred regarding the potential plasmid response to the immature sugar beet seedlings (Zhang et al., 2004a; b).

IVET technology is not without its limitations. A major limitation to investigating environmental expression in pQBR103 is that it has only been applied to investigate early plant growth stages (seedlings). pQBR103 has been demonstrated to impart a burden on *P. fluorescens* SBW25 relative to the plasmid free strain at these early growth stages (Lilley and Bailey., 1997b). Therefore the previous studies are unlikely to identify all the regions truly responsible for the environmental fitness benefits of this plasmid. However, the identification of potentially environmentally expressed regions does provide candidates for further functional analysis, as well as support that these orphan regions are transcribed.

IVET region	*dap* position	Number of CDS	CDS in region (strand)	Composition of region	Comment
1	58,998	2	054-055 (-)	1 F, 1 C	Contains *helA*§
2	80,157	1	74 (-)	1 F	Contains *helC*§
3	171,715	3	163-165 (+)	1 C, 2 O	
4	180,322	2	171-172 (-)	1 F, 1 C	
5	245,345	11	238-249 (+)	1 F, 1 C, 9 O	
6	257,946	1	265 (+)	1 O	
7	259,778	3	268-270 (+)	3 O	
8	269,256	1	283 (-)	1 O	
9	296,002	9	315-323 (-)	1 F, 1 C, 7 O	Contains *orn*±
10	322,194	32	327-363 (+)	3 F, 4 C, 25 O	Contains *helB*§
11	374,762	9	421-429 (+)	4 F, 1 C, 4 O	

Table 2.7 Potential pQBR103 CDS whose transcription may be induced in response to sugar beet phytosphere signals, according to IVET analysis. *dap* position refers to the end of the IVET insertion into the pQBR103 genome i.e. the inducible promoter is predicted to be upstream of this point. Number of CDS indicates the maximum number of CDS that could be co-transcribed (i.e are contiguous and are within 500 bp of each other) depending on the promoter position. Composition of the region refers to the number of CDS annotated as either functional (F), conserved hypothetical (C) or orphan (O). ± § Further characterisation of these genes are presented in Zhang et al., 2004a; b respectively.

2.4.4 pQBR103's genome size is due to ecologically relevant genes

At 425 kb, pQBR103 is a large plasmid (discussed further in chapter 5). Three hypotheses can be offered to account for this size. Firstly, this size is a consequence of genetic redundancy. Secondly, it has large regions of non-coding DNA ("junk" DNA). Thirdly, pQBR103 size reflects large amounts of functional DNA that is of ecological and adaptive benefit to host.

As introduced in chapter 1, genetic and functional redundancy is a feature of both bacterial chromosomes and plasmids isolated from soil and phytosphere environments. Analysis of pQBR103 has shown that there is evidence for genetic and potential functional redundancy, akin to what has been previously been observed, and is likely a reflection of plasmid genome innovation (chapter 1). In total, 19 homologous families (involving 51 CDS) have been identified in the pQBR103 genome (Table 2.8). These represented both paralogues (arising from gene duplication) and xenologues (homologues arising from the introduction of DNA by HGT) (Staton, 2002). Although a degree of genetic redundancy was observed for pQBR103, this was not extensive and inadequate to explain the plasmids large genome size.

Cluster	Possible cluster function	CDS number	Mixed
1	ParB partition	113, 455, 477	
2	Pilin-associated	36, 155	
3	Arylsulfatase regulator-like	**37, 38, 39, 42,** 47, **50, 51***	F,F,C,C,C,C,F
4	Twitching motility protein	32, 52	
5	DNA-binding domain protein	97, 221, 223, 337	F,C,C,C
6	Response regulator	103, 213, 421, 471, 475	
7	Transmembrane thiol:disulfide interchange protein	119, 131	
8	Transmembrane	126, 129	C,F
9	Transcriptional regulator	344, 347	F,O
10	RNA polymerase sigma factor	350, 465	
11	Plasmid replication protein	383, 445	
12	Conserved hypothetical	**43, 44, 46***	
13	Conserved hypothetical	391, 416	
14	Conserved hypothetical/Orphan	240, 375	C,O
15	Conserved hypothetical/Orphan	**327, 328**	C,O
16	Orphan	125, 156	
17	Orphan	216, 244, 247, 265	
18	Orphan	**217, 218**	
19	Orphan	233, 238	

Table 2.8. CDS clusters formed using TribeMCL. At an inflation level above 3.0 no clusters are formed; at a value of 2.0 six clusters were formed; at 1.2 19 clusters were formed. Contiguous CDS in a cluster are marked in bold. If there are functional genes in a cluster possible cluster function is ascribed by those functional genes. If clusters are a mix of CDS categories they are marked as mixed (F= functional; C= conserved hypothetical; O = orphan). * Clusters 3 and 13 are linked at low homology.

The second hypothesis that pQBR103 contains regions of junk DNA is also unlikely to explain genome size; the pQBR103 genome sequence was predicted to be 83.4% coding, with the largest gap between CDSs being 754 bp on the same strand and 986 bp on opposing strands. However, one of the main observations of the annotation of the pQBR103 sequence was the proportion of orphan sequences; 59% of the predicted CDS had no significant homology to any sequences in the public database. The identification of orphans and genes only homologous to other hypothetical proteins is common place in genome annotations (as outlined in chapter one). This was reflected in the first 330 bacterial geneomes sequenced; where analysis of the chromosomes finds that the percentage of orphans varies from 0% in *B. aphidicola* strain Bp to 46% in *Rhodopirellula balitca* SH1 (www.genomics.ceh.ac.uk/orphan_mine/orphan_home.php). Two views are taken on the identification of orphan genes; firstly, they encode novel genes that are hitherto undescribed and probably niche specific, or secondly they are the result of mis and overzealous annotation (Skovgaard et al., 2001). Based on sequence analysis and previous experimentation it is possible to give evidence that the former view is most likely true for pQBR103.

The orphan genes identified in pQBR103 were on average smaller than those designated as functional or conserved hypothetical ($p<0.001$) (Table 2.4), suggesting caution must be given to those that were particularly small in size. However, the orphans did not significantly differ in amino acid usage, G+C content or degree of sequence low complexity (all features which have been implicated in mis-annotation). Further support that these genes are functional comes from plasmid theory and fitness experiments of plasmids *in vivo*. Plasmid studies, including those on pQBR103 (Lilley and Bailey, 1997b), have shown in the absence of traits of fitness benefit, plasmid carriage exerts a burden on its host. Therefore, theory would dictate that DNA that did not at least offer periodic benefit to host i.e. "junk" would be lost to minimise this burden. For pQBR103 the temporal benefit it confers to host has been demonstrated in planta (Lilley and Bailey, 1997b). Moreover, IVET analysis has demonstrated transcription in response to environmental stimuli. Therefore, it can be proposed the large genome size is reflective of genes that have hitherto been unidentified and that these genes encode functions that are of ecological benefit to hosts in the phytosphere.

2.4.5 PCR surveys: pQBR plasmid diversity

Two PCR surveys, designed on the pQBR103 sequence, were conducted to investigate the diversity of 8 pQBR plasmids (Table 2.1), representing the three pQBR groups that have been repeatedly isolated over successive sugar beet growing seasons (group I, III and IV) (Lilley et al., 1996). In total each plasmid was tested for the presence of 35 loci, which only allows a crude assessment of diversity. However, the results suggest that there is no homology between pQBR103 (a group I plasmid) and the pQBR groups III and IV, other than mercurial resistance determinants. Sequencing the PCR products amplified from the *merA* and *merR* plasmids, revealed no sequence divergence.

The possibility that pQBR103 shares little sequence homology with other pQBR groups is an intriguing finding, (this is investigated more thoroughly in chapter 3). However, the most notable observation from these PCR surveys is pQBR group I diversity. pQBR103 was originally estimated to be 330 kb in size by RFLP analysis. pQBR4, pQBR41 and pQBR42 by the same analysis were predicted to be of similar sizes (ranging from 301-367 kb). Based on the 35 loci none of these plasmids could be distinguished from pQBR103 (Table 2.9). Furthermore, sequencing and comparing the PCR products from five loci (*merA*, *merR*, *int023* CDS 55 and CDS 371) revealed no sequence divergence between these plasmids and pQBR103. Therefore, it may be reasonable to infer these plasmids are largely synonymous.

In contrast, the smaller plasmids, pQBR44 and pQBR47, estimated as being 130 kb and 256 kb respectively, could be distinguished from pQBR103, based on the PCR surveys. A noteworthy observation from the results was that all the loci determined to be present were contiguous, as were all the loci determined to be absent (as shown in Table 2.9 and Figure 2.6). With only 35 loci, assessment of diversity was crude, however these finding suggest firstly, that there may exist a common group I region, and secondly pQBR44 and pQBR47 may be subsets of pQBR103. Indeed sequencing and comparing the three and five loci respectively that pQBR44 and pQBR47 have in common with pQBR103 revealed no sequence divergence (loci sequenced: *merA*, *merR* and CDS 0371 in both plasmids, and additionally int023 and CDS 055 in pQBR47), which fits with the hypothesis that these are subsets of pQBR103. If this is the case, this is intriguing as many of the maintenance functions and genes of interest do not lie within the region common to pQBR44, pQBR47 and pQBR103 (as shown in Figure 2.6). For example, it is known that all the pQBR plasmids are self conjugative yet the potential regions responsible for this transfer phenotype (Region number 2 and 4 on Figure 2.6 referred to as candidate regions 1 and 3 respectively in section 2.4.2.1.2) are not contained in this region. This may suggest that the transfer mechanism is not common or alternatively that the candidate transfer regions are not involved in transfer and as of yet unidentified transfer system lies in the common region.

pQBR number		4	41	42	44	47	29	55	57
RFLP Group*		I	I	I	I	I	III	III	IV
Estimated Size (kb)*		321	301	367	130	256	174	149	261
Year isolated		91	92	92	92	92	91	92	92
CDS	Annotation								
27	**Orphan**	+	+	+	-	+	-	-	-
28	**Type II/III secretion system pilus protein**	+	+	+	-	+	-	-	-
29	**Orphan**	+	+	+	-	+	-	-	-
30	**Orphan**	+	+	+	-	+	-	-	-
31	**Orphan**	+	+	+	-	+	-	-	-
32	**Pilus assembly-related protein**	+	+	+	-	+	-	-	-
33	**Pilus biogenesis-like protein**	+	+	+	-	+	-	-	-
34	**Transmembrane secretion-related protein**	+	+	+	-	+	-	-	-
35	**Transmembrane pilus-related protein**	+	+	+	-	+	-	-	-
36	**Transmembrane secretion-related protein**	+	+	+	-	+	-	-	-
int023†	Between CDS053-54, overlapping 54	+	+	+	-	+	-	-	
55†	Plant-inducible DNA helicase, HelA	+	+	+	-	+	-	-	
62	Conserved hypothetical	+	+	+	-	+	-	-	
144	Orphan	+	+	+	-	-	-	-	
151	Nucleoid-associated protein	+	+	+	-	-	-	-	
156	Orphan	+	+	+	-	-	-	-	
int730	Between CDS215-216, overlapping 216	+	+	+	-	-	-	-	
249	Restriction enzyme-related protein	+	+	+	-	-	-	-	
int036	Between CDS282-283, overlapping 283	+	+	+	-	-	-	-	
286	Conserved hypothetical	+	+	+	-	-	-	-	
297	Orphan	+	+	+	-	-	-	-	
310	Orphan	+	+	+	-	-	-	-	
318	Plant-inducible oligoribonuclease, Orn	+	+	+	-	+	-	-	
361	Plant-inducible DNA helicase, HelB	+	+	+	+	+	-	-	
371†	Helicase	+	+	+	+	+	-	-	
391	Conserved hypothetical	+	+	+	+	+	-	-	
431†	*Tn*5042-like Mercuric ion reductase, MerA	+	+	+	+	+	+	+	+
435†	Tn5042-like Mer activator/repressor, MerR	+	+	+	+	+	+	+	+
int034	Between CDS453-454, overlapping 454	+	+	+	+	+	-	-	
470	**CheB like methylesterase**	+	+	+	+	+	-	-	-
471	**Two-component regulator. Histidine kinase fused response regulator protein**	+	+	+	+	+	-	-	-
472	**Chemotaxis signalling protein, CheR like methytransferase**	+	+	+	+	+	-	-	-
473	**Methyl-accepting chemotaxis transducer protein**	+	+	+	+	+	-	-	-
474	**Chemotaxis signal transduction protein CheW**	+	+	+	+	+	-	-	-
475	**Two-component response regulator, CheY like domain**	+	+	+	+	+	-	-	-

Table 2.9. Results of the pQBR PCR surveys to detect shared loci. Symbol + indicates positive amplification of a product of appropriate size. (Primers are listed in Tables 2.2a and 2.2b). Symbol – indicates no amplification. Open rows correspond to the PCR survey designed on a partial pQBR103 sequence to investigate pQBR genetic diversity; Shaded rows refer to the PCR survey directed at the two targeted regions in pQBR103 (Section 2.3.10). pQBR57 was not included in the first survey. It has since been surveyed for the presence of *merA* and *merR*. Appropriate controls were included in both surveys (positive: pQBR103, negative: plasmid free *P. putida* UWC1 and *P. fluorescens* SBW25). *Estimated size and pQBR group is based on RFLP analysis (Lilley et al., 1996, Lilley and Bailey 1997a). † All PCR products amplified from the pQBR plasmids for these loci were subjected to DNA sequencing and comparison.

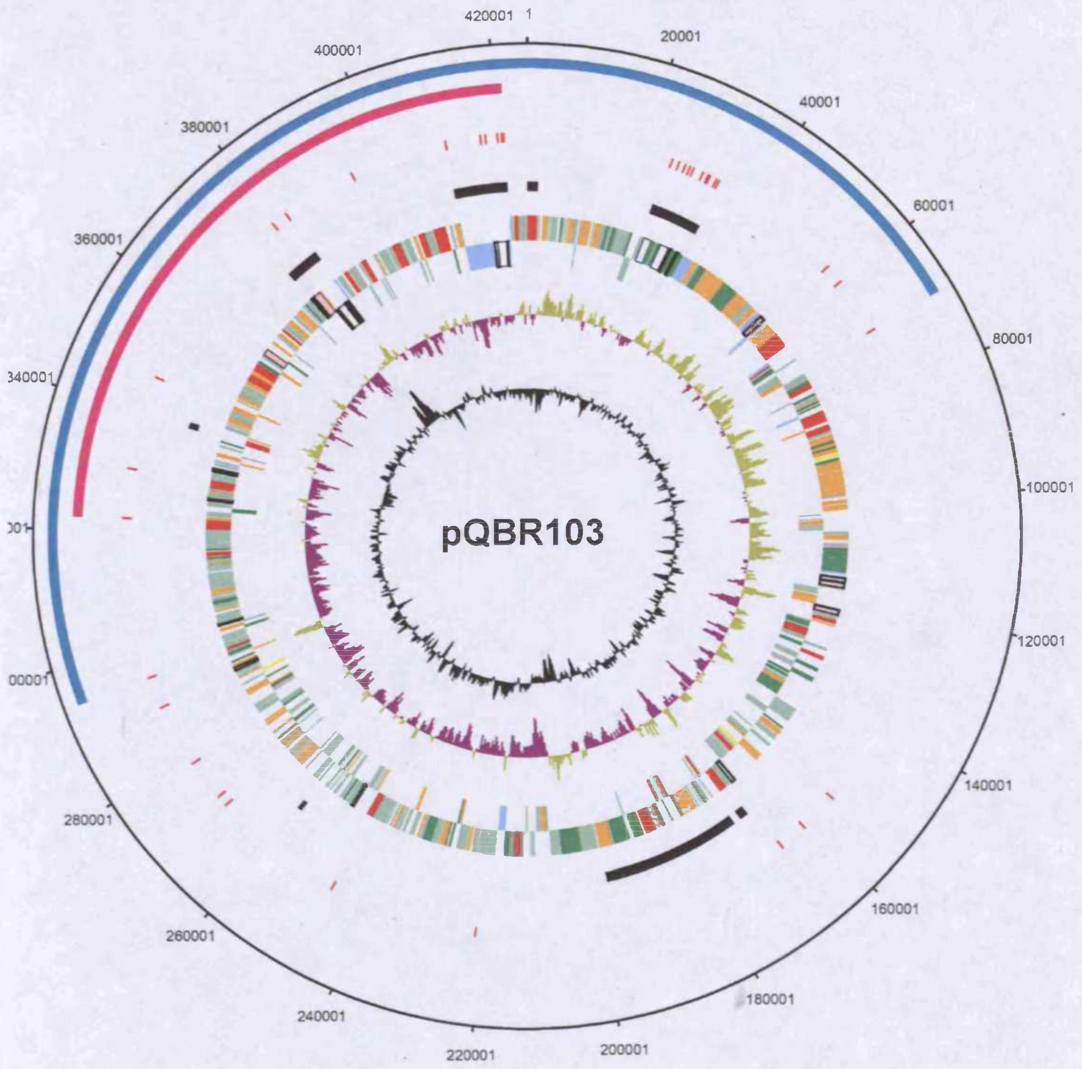


Figure 2.6 Results of pQBR PCR survey suggest their may be syntenic regions of commonality between the group I pQBR plasmids. Outermost circle, **Circle 1** (Black) represents the pQBR103 sequence, numbers refer to sequence position (bp). **Circle 2** (blue) potential region shared between pQBR103 and pQBR47. **Circle 3** (pink) potential region shared between pQBR103 and pQBR44. **Circle 4** (red blocks) the 35 loci in pQBR103 amplified by the PCR surveys. **Circle 5** (Black blocks) regions on pQBR103 of maintenance function or other interest (clockwise from pQBR103 sequence position 1) 1, plasmid partitioning region; 2, region of transmembrane /secretion homologues; 3, UV resistance genes; 4, region covering Type IV transfer like homologues; 5, region of homology to *oriV* of pQBR11; 6, replication initiation protein; 7, Tn*5042* like transposon and mercury resistance operon; 8, region of chemotaxis like homologues. **Circle 6 and 7** (multicoloured) CDS coded on forward or reverse strand respectively. Colours refer to annotated function of the gene (listed in Figure 2.1) **Circle 8** (gold and purple) G+C skew, **Circle 9** (black) GC deviation.

2.4.6 Concluding remarks

In obtaining, annotating and analysing pQBR103, insights have been gained regarding the maintenance features of this plasmid. As described for other large plasmids of low copy number, annotation has revealed that rather than relying on random segregation at host cell division, pQBR103 likely uses an active partitioning system to ensure stable inheritance. In addition to this system, pQBR103 also encodes putative proteins that may be responsible for resolving the deleterious effects of dimer formation. However, no post segregation killing system was identified. Further to these findings the probable mechanisms of replication and copy control were identified, thought to be largely similar to the iteron containing theta replicons that have been previously described (Llanes et al., 1996; Espinosa et al., 2000).

In addition to maintenance features, a number of putative traits of ecological relevance to hosts in the phytosphere were identified. These included both inorganic and organic (broad spectrum) mercury resistance, the later of which was assayed and confirmed empirically; The identification of a number of putative regulators (including a carbon storage regulator homologue and two RNA polymerase sigma factors) that may be involved in regulating global cellular responses to the environment; genes implicated in chemotaxis; and resistance to UV radiation, a trait that has been confirmed experimentally (Zhang et al., 2004b). While sequencing pQBR103 has been informative little has been added to what was already known for this plasmid and, therefore, has done little to further the understanding of this ecologically important plasmid. One of the most significant findings of sequencing pQBR103 was the proportion of the CDS that could not be ascribed a function (80%). Based on previous investigation and plasmid theory, it can be predicted that these encode hitherto undescribed genes of environmental relevance. The large proportion of the genome lacking homology to sequences in the public databases may reflect how unusual it was to sequence plasmids from natural environments (this is considered further in chapter 5). One significant finding that implies how enigmatic this plasmid is, was the absence of a characteristic conjugative plasmid transfer system, especially as self transfer is a known phenotype of this plasmid. Notably, based on the inferences of a PCR survey, the low homology candidate regions that might contribute to transfer may not be common to the group I plasmids. Indeed the PCR survey indicates that several of the annotated features may not be common including the probable genes responsible for UV resistance and an *oriV* identified in pQBR11 (Viegas et al., 1997) (this is considered further in chapter 3). Another notable finding was that there is little genetic homology between the pQBR group I and other pQBR groups, again this is investigated further in chapter 3.

What was clear from the annotation of pQBR103 was the low level of functional inferences and high levels of genes without ascribed function. Therefore, the only way forward is to adopt additional studies to further identify and assess the environmental relevance of this plasmid. In chapter 4, one such study is used to investigate pQBR103 gene expression in response to environment.

Chapter 3: Plasmid Comparative Genomic Hybridisations

3.1 Chapter Introduction

In chapter 2 PCR surveys of the pQBR plasmids demonstrated that there is genetic diversity within the group I pQBR plasmids. With the exception of the mercury resistance operon, no evidence of similarity was identified between the group I and two other pQBR groups tested (groups III/IV). From this study, of particular interest were plasmids pQBR44 and pQBR47 as the survey data revealed deletions in these but also likely syntenic regions of commonality between these and other group I plasmids. This raises the possibility that pQBR group I plasmids have a modulistic makeup, with a common "core" or "backbone". This has been observed and studied in plasmid groups, in particular for the IncP catabolic plasmids (Dennis, 2005) and the IncH plasmids (Gilmour et al., 2004).

In the previous chapter "probes" for the PCR surveys were initially designed on a partial genome sequence (over 100 contigs of 2-30 kb) due to a partially successful in house sequencing effort. Since the complete nucleotide sequence of the plasmid has been determined it is possible to use this sequence to create a "genomotyping" array (Lucchini et al., 2001). Although such an array is unable to replace the benefit of obtaining complete sequences of many closely related strains or isolates, it will enable further investigation of the genetic composition of the pQBR plasmids at a higher resolution than that afforded by the PCR surveys (chapter 2). The usefulness of genomic array approaches in the study of genomic compositions has been demonstrated previously (Salama et al., 2000; Zhou et al., 2004; Zhao et al., 2005).

DNA arrays designed for genomotyping vary considerably, but all harness the same basic principle of hybridisation between complementary sequences. Essentially there are two major differences in genomic array design. These are the type of DNA probe used and the physical support for those probes (introduced in chapter 1). Usually probes are designed for each CDS of the genome and are either chemically synthesised or amplified by PCR. However, a cheap and easy alternative in probe construction is the amplification of inserts from the clone library originally created for genomic sequencing. The first example of this was the *Campylobacter jejuni* comparative genomic hybridisation (CGH) array (Dorrell et al., 2001). In respect to the choice of probe support there are essentially two types, nylon or glass/silicon. Although nylon microarrays do exist (Honore et al., 2006), generally these arrays are larger and as such referred to as macroarrays. In comparison glass/silicon are smaller and termed microarrays. There are differing advantages and limitations to both macro and microarray approaches, but both have been used successfully in comparative analysis of both bacterial and plasmid genomes (Dorrell et al., 2001; Zhao et al., 2005). In this chapter both array approaches, macro and micro, will be assessed for utility. Ultimately one will be employed to investigate the intra group dynamics of the pQBR group I plasmids, as well as their relatedness of other pQBR groups that are persistent at the Oxford Wytham field site (group III and IV) (Lilley et al., 1996).

3.2 Chapter Aims

In this chapter the aims are

1) To investigate the feasibility of using different array approaches i.e. nylon macroarrays and glass microarrays in comparative plasmid analysis.
2) On selection of the technology, create a "genomotyping" array based on probes amplified from the pQBR103 sequence clone library.
3) Challenge the array with pQBR group I representatives.
4) Challenge the array with non pQBR group I representatives (from groups III and IV).

This will address the following hypotheses/questions

1) It is hypothesised that there exists a group I "core" region. Firstly, what are the boundaries of this core and are there genomic features flanking it to suggest mechanisms of insertions or deletion? Secondly what plasmid encoded traits are conserved in this core and how does this compare to the designated backbones described for other plasmid groups (Dennis, 2005: Gilmour et al., 2004)?

2) It is hypothesised the group I "core" is not shared by either group III or group IV. Finer scale assessments will address the question what is the extent of the genetic commonality between pQBR103 and the group III and IV plasmids? More specifically

is there any evidence to suggest intergroup genetic exchange or conservation of maintenance functions with pQBR103 to indicate a common ancestral origin?

3.3 Materials and methods

All reagents used were from Sigma-Aldrich unless otherwise stated.

3.3.1 Bacterial culture conditions and DNA extraction

The strains and plasmids used in the macroarray and the microarray experiments are listed in Table 3.1. All strains were recovered from glycerol freezer stocks and spread onto PSA plates with 27 μg ml^{-1} $HgCl_2$, where appropriate, and incubated at 28 °C for 48 hours.

All strain/plasmid DNAs were extracted using the CTAB method described previously (section 2.3.2, chapter 2) apart from the test DNAs applied to the microarray which were extracted using the method of McAllister and Stephens, 1993.

3.3.1.1 The McAllister and Stephens extraction method

Briefly, cells from each growth plate were resuspended in 15 ml of lysing buffer (10 mM NaCl, 20 mM Tris-HCL pH 8.0, 1 mM EDTA, 100 μg ml^{-1} proteinase K, 0.5% SDS) mixed and incubated for 6 hours at 50 °C. After incubation 10 ml phenol/chloroform (50:50) was added to each sample, agitated for 10 minutes at room temperature and centrifuged for 20 minutes at 1,400 x *g* before the supernatant was transferred to a clean tube and centrifuged for a further 10 minutes. DNA was precipitated by the addition of 1 ml 3 M sodium acetate pH 5.2 and two volumes ethanol, spooled using a glass Pasteur pipette (Volac, Cole Parmer, UK) and the precipitate washed twice in 70% ethanol spray before finally being resuspended in 2 ml 10 mM Tris-HCL.

Bacterial strains and plasmids	Name	Plasmid Group[§]	Estimated plasmid size (kb)[§]	Used in Macroarray analysis	Used in Microarray analysis
Plasmid free Pseudomonad strains	*P. putida* KT2440	-	-	yes	no
	P. putida UWC1	-	-	yes	yes
	P. fluorescens SBW25	-	-	yes	no
Plasmids hosted in *P. putida* UWC1	pQBR4	I	321	yes	no
	pQBR41	I	301	yes	no
	pQBR42	I	367	yes	no
	pQBR44	I	130	yes	yes
	pQBR47	I	256	yes	yes
	pQBR29	III	174	yes	no
	pQBR55	III	149	yes	yes
	pQBR57	IV	261	no	yes
	pQBR103	I	330	yes	yes
Other	*E. coli* DH10	-	-	no	yes

Table 3.1. Pseudomonas stains and plasmids used in chapter 3. § Groups and size estimated by RFLP analysis described in Lilley et al., 1996. *E. coli* from Invitrogen, UK. All other strain and plasmid references are given in Table 2.1 (chapter 2).

3.3.2 Macroarray Analysis

3.3.2.1 Macroarray probe design, amplification and quality control

The 19 probes used represented 250 bp regions of the pQBR103 genome, designed and amplified by the primers that are listed in chapter 2 (section 2.3.10, Table 2.2). To avoid confusion when referring to "probe" in the context of the macroarray experiments, this represents the 19 PCR products regardless of whether they are immobilised on the membrane or labelled and challenge to the array.

The amplification of the probes used the primers, the PCR reagents and cycling conditions, previously described in chapter 2 (section 2.3.3.1). All PCR reactions were visualised and product sizes and purity confirmed by gel electrophoresis. Products were excised from the gel and purified using the Qiagen gel extraction kit (Qiagen, UK) according to manufacturer's instructions.

3.3.2.2 Macroarray design and construction

Two differing macroarray designs were used to test feasibility of the method.

1) The 19 PCR probes were spotted onto the nylon support in duplicate.

2) Purified DNA from the plasmid free pseudomonad strains plus the 8 pQBR plasmid in pseudomonad hosts were immobilised in duplicate on the nylon support.

This method of macroarray construction using a suction manifold was based on that previously described (Brown, 1994-2005). Firstly, a positively charged nylon membrane (Hybond-N+ Amersham Pharmacia Biotech, UK) was cut to the size of the manifold (Hybri.dot manifold, BRL life technologies, USA) and wetted for 10 minutes in milliQ, before being placed in the manifold over a filter support (Whatman^{TM} 3MM). The membrane was prepared for spotting by prewashing 500 μl of milliQ water through each of the wells in the manifold.

Sample DNAs* to be immobilised to the nylon membrane were prepared by adding 1 M NaOH and 200 mM EDTA pH 8.2 to give a final volume of 350 μl and final concentration of 0.4 M NaOH, 10 mM EDTA.

* 500 ng of total probe DNA was applied per spot in the construction of macroarrays with immobilised probes (design 1). Additionally 0.5 μg of pQBR103 was applied. In array design 2 (host/plasmid DNA immobilised onto the membrane), 3.5 μg of each genomic extraction was applied per spot.

Samples were denatured by incubating in a water bath at 100 °C for 5 minutes before being applied to the membrane wells. Once the samples had filtered through the membrane, 500 μl of 0.4M NaOH post wash was washed through. The membrane was removed from the manifold and briefly washed in 2 x Standard Saline Citrate (SSC) (Standard 1 x concentration: 0.15 M NaCl, 0.015 Na citrate).

These concentrations and detection limits were theoretically and empirically estimated using control DNAs. Initially, different amounts of control lambda phage DNA (0.5, 5, 50, 500, 5000 pg) supplied by manufacturer (Amersham Pharmacia Biotech, UK) were spotted onto the membrane and challenged with self, using the hybridisation conditions outlined below. Secondly, a macroarray was constructed by spotting 350 ng and 3.5 μg spots of *P. putida* UWC1 containing pQBR55 and *P. fluorescens* SBW25 containing pQBR103 to the membrane as well as varying amounts of the *mer*R probe (50 pg, 500 pg and 5 ng). This array was then challenged with a *mer*R probe using the hybridisation protocol given below.

3.3.2.3 Macroarray Hybridisations

Sample DNA was labelled using the commercial enhanced chemiluminescence (ECL) labelling system of Amersham Pharmacia Biosciences, UK (RPN3000). Protocols were largely based on manufacturer's recommendations.

3.3.2.3.1 DNA labelling

In both macroarray designs 10 µl of 20 ng µl^{-1} host/plasmid DNA or PCR probe was labelled. Firstly DNA was denatured by boiling at 100 °C for 5 minutes before being quickly cooled on ice. 10 µl of labelling reagent (containing peroxidase) was added. To covalently link the peroxidase to the DNA 10 µl of glutaldehyde was added and left to incubate for 10 minutes at 37 °C.

3.3.2.3.2 Hybridisation conditions

To the supplied hybridisation buffer base, solid NaCl and blocking solution was added and dissolved to give a final concentration of 0.5 M NaCl, 5% (w/v) blocking reagent. Each nylon membrane was prepared by rinsing in 5 x SSC before being unrolled in a hybridisation container. Buffer preheated to 42 °C was added to the container and agitated in a hybridisation rotisserie oven (Hybridiser HB-ID, Techne, UK) for 15 minutes prior to the addition of the labelled probe. Hybridisation was continued overnight (~16 hours) with gentle agitation at 42 °C.

3.3.2.3.3 Array washing, signal detection and interpretation

After the required time for hybridisation had elapsed (overnight) the buffer was removed from the hybridisation bottle and replaced with 50 ml 5 x SSC and agitated for 5 minutes before discarding. Primary wash buffer (6 M UREA, 0.4% SDS, 0.5 x SSC) was preheated to 42 °C and 100 ml was added to the container and agitated for 20 minutes at 42 °C. Primary wash buffer was discarded and an equal volume replaced and allowed to incubate for 10 minutes. This wash was repeated twice. Arrays were then removed to an excess of secondary wash buffer (2 x SSC) and agitated on an orbital shaker for 5 minutes before being repeated. Secondary wash was drained from the arrays. To detect successful hybridisation an equal volume of detection reagent 1 was added to reagent 2 and poured evenly over the arrayed surface of the membrane, and incubated at room temperature for 1 minute. Excess detection fluid was drained from the arrays and luminescence detected by exposing the array for 1hour in a BioRad Versa

Doc system using BioRad software (BioRad, UK) at default settings according to the manufacturers instructions.

Only qualitative data was captured for each hybridisation i.e. each probe was only recorded as positive or negative. This assessment was based on the authors interpretation.

3.3.3 Microarray Analysis

3.3.3.1 Microarray probe design, preparation and quality control

Non overlapping clones were selected from the pQBR103 pUC19 sequencing library that spanned the entire length of the plasmid genome. The length of sequence unrepresented between adjacent clones was kept to a minimum. These inserts constituted the probes for the pQBR103 custom microarray.

To detect probes that might cause cross hybridisation with host or non specific pQBR103 regions of the pQBR103 genome, each of the selected probes was queried using the blastn (Altschul et al., 1997) (version 2.2.12 matrix = 1-3, gap existence 5, extension = 2) against the full nucleotide sequence of *P. putida* KT2440 (Nelson et al., 2002) and pQBR103. Probes showing 80% nucleotide similarity over 30 nucleotides were flagged as probes that could potentially show cross hybridization with host chromosomal DNA or to homologous plasmid regions.

Construction of probes was by PCR amplification of the selected pUC19 clones using the DNA extractions originally performed for genome sequencing as template for the reactions (chapter 2).

All PCRs were performed using the following primers (note, primers are given in 5' to 3' orientation): M13f; TGTAAAACGACGGCCAGT and pUCR; GCGGATAACAATTTCACACAGGA. All PCRs were performed using BioLine (London, UK) reagents and the following reaction conditions for 50 µl volumes: Template DNA ~50-100ng (1 µl), Each primer 10 pMol stock (1 µl), 10 x PCR buffer

(5 μl), dNTP 25 mM (1 μl), $MgCl_2$ 50 mM (1.5 μl) Taq polymerase 5 U μl^{-1} : PFU 2.5 U μl^{-1} mix (8:2, v/v) (0.5 μl), milliQ (40 μl). Cycling conditions: Initial denaturation 3 minutes 94 °C, then (1 minute 94 °C, 1 minute 54 °C, 7 minute 72 °C) for 30 cycles, Final extension 10 min 72 °C, hold at 4 °C. Cycling reactions were conducted on a PTC-225 DNA engine tetrad (MJ Research, UK).

PCR reactions of each clone insert were repeated in triplicate. Positive reactions were confirmed via agarose gel electrophoresis. Negative amplifications were repeated a further two times, as were amplifications showing non-specific primer binding, before the probe was excluded. Successful amplifications were pooled for each clone before being purified through Sigma PCR purification columns as per manufacturers' instructions (Sigma-Aldrich, UK). Purity, concentration and probe size were estimated by agarose gel electrophoresis using markers of known DNA size and amounts (Bioline Molecular ladder 1). All probes were either diluted or concentrated to approximately 100 ng μl^{-1}. In addition to determining probe size, a subset of 10 probes were sequenced using ABI technology as quality control to ensure the correct amplification of clone inserts.

3.3.3.2 Microarray design and construction

Slides were spotted by Dr Nigel Saunders of the Dunn School of Pathology, University of Oxford.

Slide construction: Each probe was suspended 50:50 in Genetix spotting solution for amine slides and printed onto aminosilane slides using a Genetix Qarray mini microarray printer with solid tungsten 150 mm aQu pins. After printing the slides were kept humidified for 12 hours, baked for 30 minutes at 85 °C and UV irradiated with 300 J. Circularity and uniformity of spots was checked by scanning each printed slide at high intensity at 550 nm using a GenPix 4000 system (Molecular Devices, UK).

In total 122 pQBR103 probes were spotted 6 times per slide in a nonsequential manner. In addition pure pUC19 DNA (Invitrogen, UK), total *E. coli* DH10 DNA (Invitrogen, UK) and total *P. putida* UWC1 were also printed on the array representing positive and negative controls. Each slide was divided into two subarrays, A and B (Figure 3.1) which were direct repeats of each other. Each subarray was further divided into 4 arrays consisting of 32 probes repeated three times.

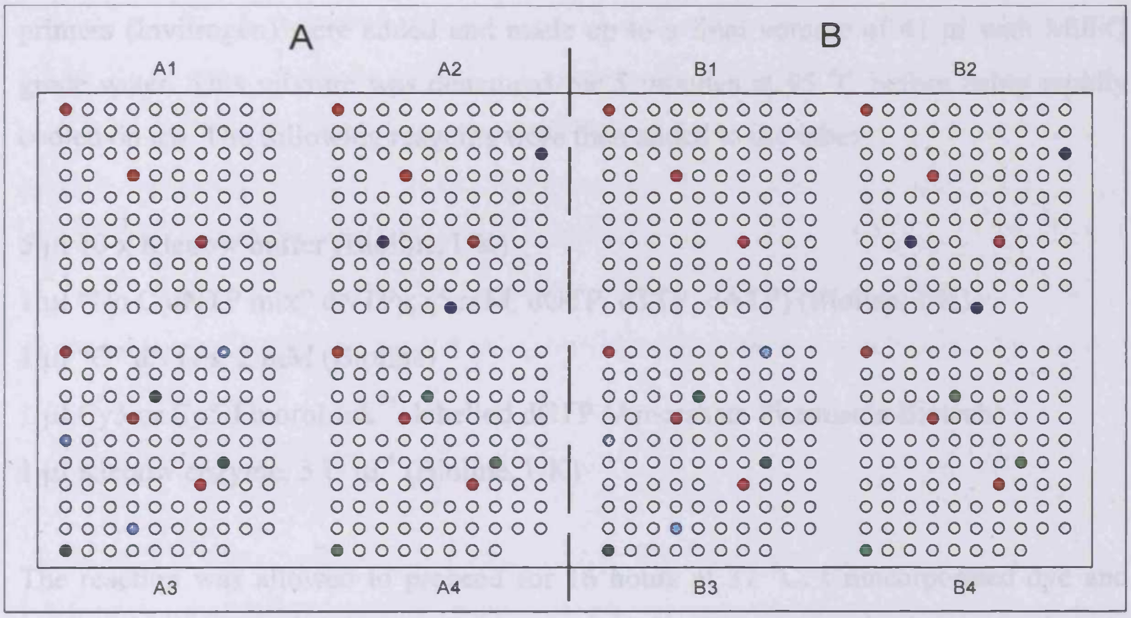


Figure 3.1. Design of pQBR103 custom microarray. Subarrays A and B are identical replicates. In total 122 test probes (including the two probes containing regions of the mercury operon shown as light and dark blue spots), 1 pUC19 probe (dark green spots) and 1 *E.coli* DH10 DNA probe (light green spots) were replicated 3 times per subarray (6 spot replicates in total per slide). *P. putida* UWC1 DNA (red spots) were spotted 12 times per subarray (24 spot replicates per slide).

3.3.3.3 Microarray slide preparation

Each microarray slide was prepared for hybridisation by incubating the slide for 20 minutes at 65 °C in a blocking solution comprising, 6 x SSC, 0.1% SDS, 10 mg ml^{-1} BSA (Sigma, UK). Following incubation slides were washed, firstly in MilliQ water for 1 minute, then isoproponol for a further minute before being immediately dried using an airbrush (Paasche, with SimAir compressor). To check for the introduction of smears/streaks each slide was visualised on a ScanArray Express HT microarray scanner (Perkin Elemer, UK).

3.3.3.4 Labelling reactions

Labelling reactions were based on those described by Snyder et al. (2005). The six target DNAs were directly labelled. To 20 μg of a target DNA 3 μg of random hexamer primers (Invitrogen) were added and made up to a final volume of 41 μl with MilliQ grade water. This mixture was denatured for 5 minutes at 95 °C before being rapidly cooled on ice. The following reagents were then added to the tube:-

5 μl 10 x Klenow buffer (Bioline, UK)
1 μl "No C dNTP mix" dNTPs, (5 mM, dGTP, dTTP, dATP) (Bioline, UK)
1 μl "C" dNTPs, 2 mM (Bioline)
1 μl Cy3 or Cy5 FluoroLink™ labelled dCTP (Amersham Pharmacia Biotech)
1 μl Klenow enzyme, 5 U μl^{-1} (Bioline, UK)

The reaction was allowed to proceed for 16 hours at 37 °C. Unincorporated dye and primers from each hybridisation reaction were removed by using the Qiagen nucleotide removal kit (Qiagen, UK) as per manufacturer's instructions, eluting in 50 μl MilliQ water. The amount of labelled DNA and incorporation of labelled nucleotides was determined by spectroscopy using a NanoDrop ND-1000 UV-vis system (LabTech international, UK) based on the following equations and assumptions (where A ≡ Absorbance).

Amount of target DNA (ng) = A_{260} x 37 x total volume DNA (μl)

Labelled nucleotide incorporation (pMoles)
For Cy3: A_{550} x total volume of DNA/0.15
For Cy5: A_{650} x total volume of DNA/0.25

Each labelled reaction was concentrated by evaporation to a final volume of 17.1 μl using a speed evaporator as per manufacturer's instructions (Eppendorf, UK). Labelled target was prepared for hybridisation by adding SSC and SDS to a final concentration of 4 x SSC, 0.29% SDS and final volume of 26 μl.

3.3.3.5 Slide hybridisation and fixing

Slide hybridisation and fixing was performed as described by Snyder et al. (2005). Each labelled reaction was applied to a 65 °C preheated hybridisation chamber (Corning, UK) containing a blocked microarray slide and 22 mm x 22 mm lifter slip (Erie Scientific, USA). The hybridisation chamber was kept moist by adding 15 μl hydration solution (4 x SSC, 0.3% SDS) to each humidity well before the cassette was sealed according to manufacturer's instructions. The cassettes were incubated in a water bath at 65 °C for 16 hours. After incubation the slides were washed and fixed by washing in 65 °C wash A (1 x SSC, 0.05% SDS) for 2 minutes before washing twice for two minutes in wash B (0.06 x SSC) and completely dried using an airbrush (Paasche, with SimAir compressor)

3.3.3.6 Data extraction

Slides were scanned using a ScanArray Express HT microarray scanner (Perkin Elemer, UK). Each labelling and hybridisation reaction was repeated twice for each test strain, representing independent slide replicates. Data was extracted from the slide images using GenePix Pro 6 software (Molecular Devices, UK). Hybridisation spots that were severely deformed i.e. comet tailed, or in areas of precipitate or local high intensity were excluded from further analysis. Median fluorescence intensity was extracted for each spot to correct for irregularity in spot shape. Background fluorescence was corrected for by subtracting median background intensity for each probe. A common way to globally normalise a single channel array is by dividing the intensity values for each spot by either the average intensity of some "housekeeping genes" or by the average intensity across all spots. However, because of the large variability in positive results in different arrays (~ 12 - 600) this is known not to work well and so was not appropriate (Call et al., 2006). Instead for each probe (represented 6 times on each array) a single intensity value was given as the average median for each replicate. The within array variability for each probe was given by calculating the standard error from each replicate. Based on the array hybridisation an arbitrary cut-off was given for positive results of 10,000 intensity units, a method that has been used previously for such studies (Call et al., 2006). This cut-off was determined by the control spots on the array i.e. all positives were above this level and all negative controls below.

There was no mechanism for cross array normalisation in the experimental design. Therefore, spot intensities could not directly be compared between arrays. Instead comparisons were made for each slide replicate by comparing qualitative data from each of the slide replicates i.e. whether probes determined present or absent (using the above criteria) were congruent between the two slides replicates.

3.4 Results

3.4.1 Developing a macroarray

3.4.1.1 Macroarray option 1: Probe immobilised to the membrane

The control macroarray was constructed using varying concentrations of lambda phage DNA (0.5 pg – 5 ng) to test the sensitivity of the array approach. When challenged with 100 ng of self (labelled lambda), hybridisation was only observed where probe DNA amount was 0.5 ng or greater. This corresponds to a 1000 fold lower sensitivity than that claimed by the manufacturers.

Based on the control array total amounts of 500 ng of each plasmid test probe was immobilised to the membrane. Also the amount of labelled DNA was doubled to 200 ng. This practically represented the maximum amount of probe and labelled DNA that could be used (i.e. above these amounts probe construction and DNA labelling is deemed not to be time and cost effective). Nevertheless, challenging the array with *P. fluorescens* SBW25 carrying pQBR103 total genomic DNA resulted in no detectable signal. Because of the lack of sensitivity this array option was discounted.

3.4.1.2 Macroarray option 2: Immobilising total genomic DNA to the membrane

The sensitivity of detection using this macroarray option was calibrated using a test array (macroarray spotted with 50 pg - 5 ng *merR* probe and 350 ng and 3.5 μg of *P. putida* UWC1/pQBR103 and *P. putida* UWC1/pQBR55 genomic extractions). The results of this calibration array are shown in Figure 3.2a. When challenged with the *merR* probe, hybridisation signal was observed for probe hybridising with self. Positive hybridisations were also observed for both 3.5 μg DNA spots of the pseudomonad/plasmid genomic extractions. Based on the size of the host chromosome and the assumptions that the plasmid is of low or single copy number, as well as the gene target this equates to a detection limit of around 0.5 ng which is congruent to the detection limit observed with the lambda test DNA (details above). Based on these results 3.5 μg of genomic DNA was spotted onto the membranes for each pseudomonad/plasmid combination. The hybridisation results for four macroaarys challenged with four different probes are given in Figure 3.2b.

Plasmids that showed positive hybridisations on the arrays challenged with probes int023 and int034 are congruent with the results obtained with the PCR screen (chapter 2, section 2.4.5, Table 2.9) i.e. plasmids pQBR4, pQBR41, pQBR42, pQBR47 are positive for both probes and pQBR44 is positive for int034 but negative for int023. However, converse to what was expected, positive hybridisations were observed with *P. putida* KT2240 when challenged with int034, int023, *merA*, *merR*; *P. putida* UWC1 when challenged Int034, *merA* and *merR*; and *P. fluorescens* when challenged with *merA*. These results are likely to be false positives as complete sequence information is available for these strains.

Of the two options only the second yielded results, but because of the lack of sensitivity and discriminative power of the array this work was discontinued in favour of a microarray technology.

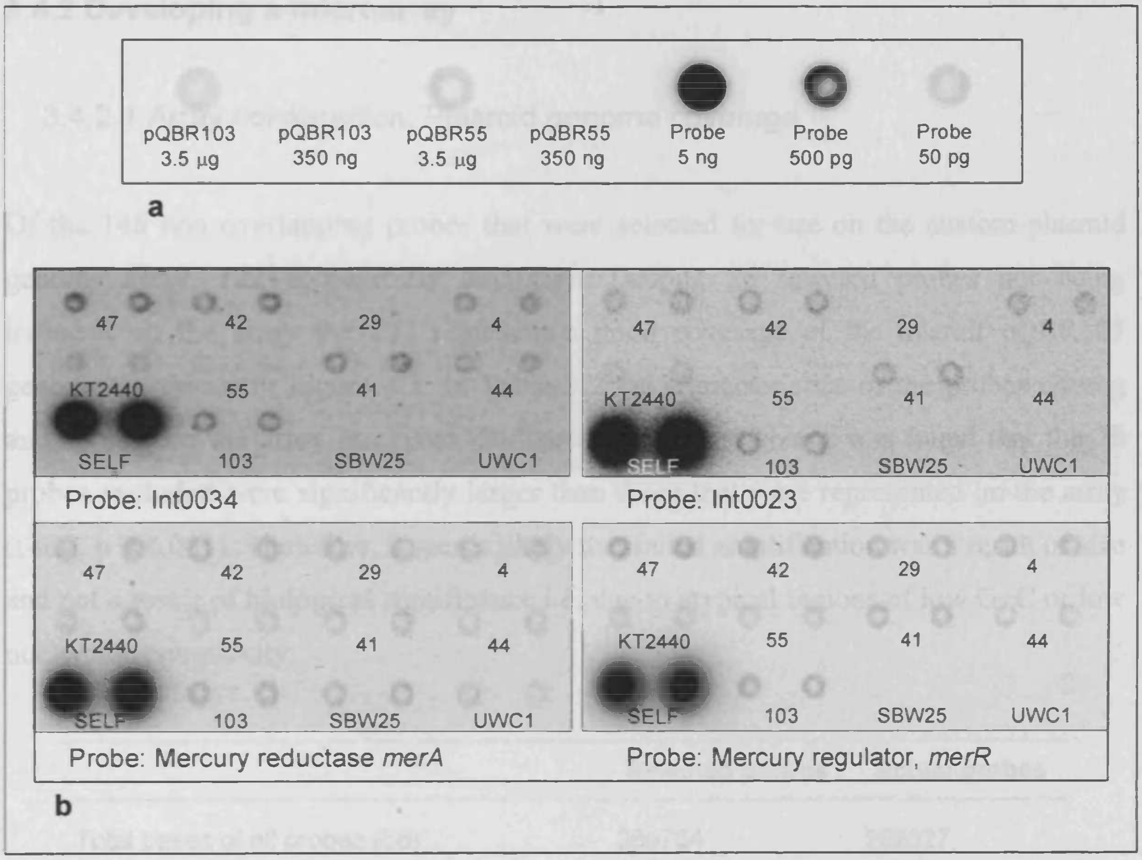


Figure 3.2., a) Test macroarray hybridisation result when challenged with a *merR* probe. Probe in Figure refers to amount of *merR* probe that has been bound to the membrane b) Four macroarray hybridisation results when *pseudomonas* DNA and *pseudomonas*/plasmid DNA was bound to the membrane and challenged with four different probes; int0034, int0023, *merA* and *merR*. Numbers refer to the pQBR plasmid number. KT2440 and UWC1 refers to the strain of *P. putida*, SBW25 refers to strain of *P. fluorescens*. Self in Figure refers to positive control where probe that challenged the array was also bound to the membrane.

3.4.2 Developing a microarray

3.4.2.1 Array construction: Plasmid genome coverage

Of the 148 non overlapping probes that were selected for use on the custom plasmid genome array, 122 successfully amplified. Despite 26 selected probes not being included on the array the 122 represent a good coverage of the overall pQBR103 genome as shown in Figure 3.3. In Table 3.2 the characteristics of the probes chosen and spotted on the array are given. By comparing probe size it was found that the 26 probes excluded were significantly larger than those that were represented on the array (t-test, $p < 0.001$). Therefore, it seems likely that failed amplification was a result of size and not a result of biological significance i.e. due to atypical regions of low G+C or low nucleotide complexity.

	Selected probes	**Actual probes**
Total bases of all probes (bp)	369754	292527
Size of pQBR103 (bp)	425094	425094
Percentage of sequence represented	87	69
Average size of probe (bp)	2498	2398
Smallest probe (bp)	1415	1415
Largest probe (bp)	3954	3388
Number of probes above 3 kb	21	9
Number of probes above 2.5 kb	69	47
Number of probes above 2.0 kb	142	117
Number of probes above 1.5 kb	146	120
Number of probes above 1 kb	148	122
Number of probes between 1-1.5 kb	2	2
Number of probes between 1.5-2 kb	4	3
Number of probes between 2-2.5 kb	73	70
Number of probs between 2.5-3 kb	48	38
Average gap size between probes (bp)	375	1088
Minimum gap length (bp)	1	1
Maximum gap length (bp)	3105	8185

Table 3.2. Summary of the distribution and characteristics of probes selected to be represented on the pQBR103 microarray in comparison to the probes that were actually represented on the array.

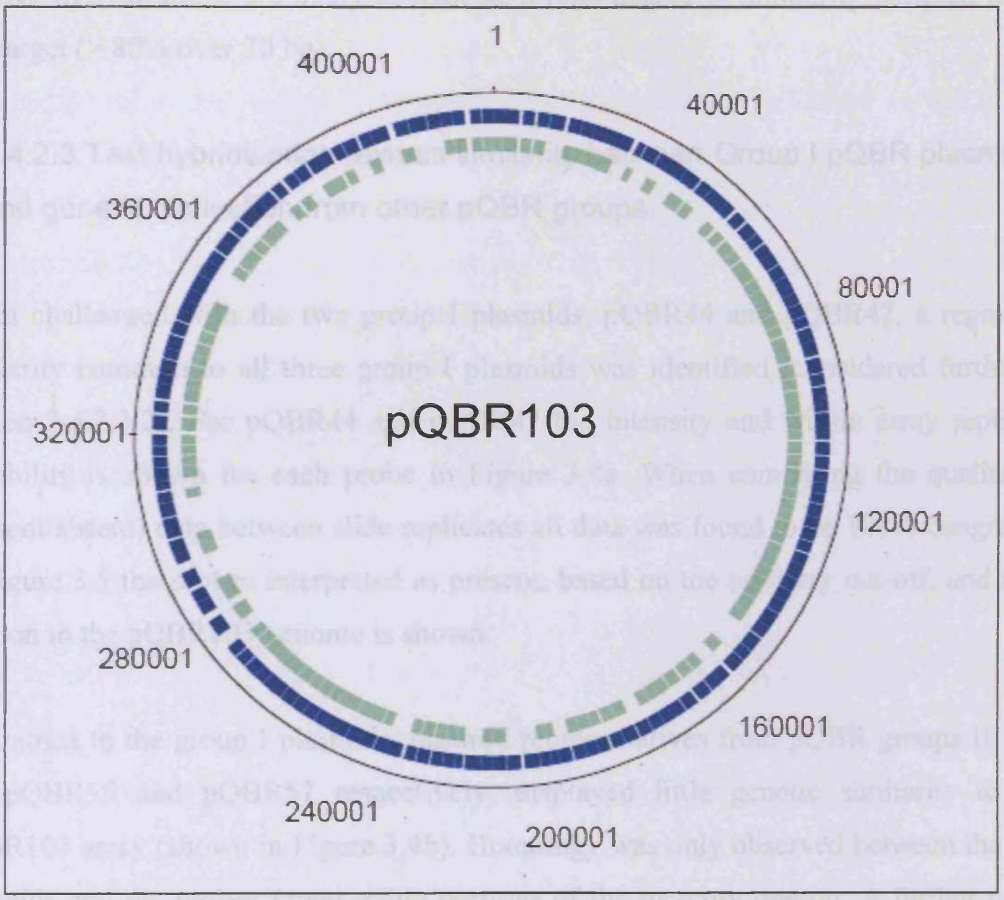


Figure 3.3. Distribution of probes selected and amplified for the custom microarray. From outermost circle to centre: **Outermost circle** (black) represents the pQBR103 sequence, numbers refer to position in that sequence (bp). **Circle 2** (blue) 148 representative probes of pQBR103 that were selected for microarray construction. **Circle 3** (green) 122 selected probes that successfully amplified and were used in array construction.

3.4.2.2 Validation of custom array: Control hybridisations

The array was challenged and validated with two controls. The positive control was pQBR103. Positive signals where detected for all probes and positive controls on the array for both slide replicates. The median probe intensity and within array variability for each probe when challenged with pQBR103 is shown in Figure 3.4a. For the negative control, *P. putida* UWC1 without plasmid, all probes were identified as negative. Negative hybridisations were also observed for probes that were flagged as sharing 80% homology over 30 nucleotides. Although, the exact stringency of the microarray cannot be determined, an assessment can be made as the level of sequence similarity between the array probes and *P. putida* UWC1 is known. Based on this, positive hybridisations are likely to indicate a high degree of similarity between probe and target (> 80% over 30 bp).

3.4.2.3 Test hybridisation reveals similarity between Group I pQBR plasmids, and genetic distinction from other pQBR groups.

When challenged with the two group I plasmids, pQBR44 and pQBR47, a region of similarity common to all three group I plasmids was identified (considered further in section 3.4.2.3.2). For pQBR44 and pQBR47 the intensity and within array replicate variability is shown for each probe in Figure 3.4a. When comparing the qualitative (present/absent) data between slide replicates all data was found to be 100% congruent. In Figure 3.5 the probes interpreted as present, based on the arbitrary cut-off, and their relation to the pQBR103 genome is shown.

In contrast to the group I plasmids, plasmid representatives from pQBR groups III and IV, pQBR55 and pQBR57 respectively, displayed little genetic similarity to the pQBR103 array (shown in Figure 3.4b). Homology was only observed between the two plasmids and the probes representing portions of the mercury operon. A further probe displayed a low level of hybridisation signal that straddled the cut-off value for interpretation. This can most likely be interpreted as a divergent sequence (This clone represents a hypothetical protein). From these results it can be inferred that pQBR55 and pQBR57 at least encode different plasmid maintenance functions (e.g. replication, partitioning) from pQBR103.

In chapter two, PCR surveys were carried out to study the diversity of the pQBR plasmids (Section 2.4.5). Plasmids relevant to this chapter, pQBR44, pQBR47, pQBR55, and pQBR103 were tested for the presence and absence of 35 loci (plasmid pQBR57 was only tested for 16 loci). Comparing the results from this chapter with the PCR survey all results are 100% congruent i.e. all loci deemed present or absent by PCR survey are similarly supported by the microarray experiments.

3.4.2.3.1 Similarity between pQBR group I plasmids

Figure 3.5 shows the regions of similarity between plasmids pQBR44, pQBR47 and pQBR103. For the most part it was only possible to give an approximation of where the region of similarity starts and ends between plasmids, because the large probe size decrease the resolving power of the array. The regions of similarity pQBR44 and pQBR47 have with pQBR103 are as follows (travelling clockwise via the 0 bp position of the pQBR103 genome):

- pQBR44 shares similarity to the pQBR103 region spanning ~300, 000 bp to ~5, 500 bp
- pQBR47 shares similarity to the pQBR103 region spanning ~290,000 bp to ~150, 000 bp

Comparing the previously determined plasmid genome sizes for pQBR44 and pQBR47 (Lilley et al., 1996) with the size of the regions they share with pQBR103, it is possible to give an estimation of the proportion each plasmid genome has in common. Table 3.3 shows these comparisons. For completeness, pQBR55 and pQBR57 have also been included. This data suggests the proportion of the genomes pQBR44 and pQBR47 shares with pQBR103 is close to 100% of their respective DNA content. Therefore they represent subsets of the pQBR103 genome, and are possible deletion variants. This, however, does not take into consideration probes representing regions outside those bulleted above that positively hybridised (shown in Figure 3.4). Control hybridisations showed the probes did not cross hybridise with host (described above). Also, all probes were tested for cross hybridisation with non specific regions of pQBR103 using *in silico* methods; therefore, theoretically these positive probes cannot be explained by cross hybridisation to host or pQBR103 sequences (other than one, see below). Thus, positive signals are likely attributable to the plasmid sequences. This suggests that even though pQBR44 and pQBR47 are obvious subsets of pQBR103, they may contain a small amount of novel DNA and/or display evidence of internal re-assortment events.

One probe could be identified as possibly showing cross hybridisation. When the array was challenged with pQBR44, a probe outside the bulleted region above showed positive hybridisation. Analysis of this probe found it to contain a paralogous CDS to one in the region shared with pQBR103, therefore potentially this signal could have been because of cross hybridisation.

	RFLP size estimate (actual size) Kb	Estimated size of homology with pQRB103 kb*	RFLP group	Reference
pQBR103	330 (425)	N/A	I	§
pQBR44	130	~127	I	±
pQBR47	256	~285	I	±
pQBR55	149	7	III	±
pQBR57	261	7	IV	±

Table 3.3. Comparison of RFLP estimated size and the length of the shared region with pQBR103. * Size includes only consecutively positive probes (shared region given in text). Single positive hybridisations are excluded in size estimate. ± Lilley and Bailey, 1997a. § Lilley et al., 1996.

Both the genomes of pQBR44 and pQBR47 can be explained as deriving from pQBR103 by deletion events. Of course an alternative explanation is that pQBR103 has arisen from insertions into pQBR44 and or pQBR47. Due to the resolution of the array the specific areas on pQBR103 where insertions or deletion events (indels) have occurred (flanking the regions homologous to pQBR44 and pQBR47) can only be approximated. These regions are herein called the breakpoints. Analyses of these breakpoints have identified no specific mechanism for insertion or deletion i.e. there were no obvious repetitive elements and insertion sequences or regions of significant homology (> 15 bp). Also no G+C deviation or change in strand bias was detected. (Although there was some G+C skew associated with the region common to pQBR47 and pQBR103, shown in Figure 3.5)

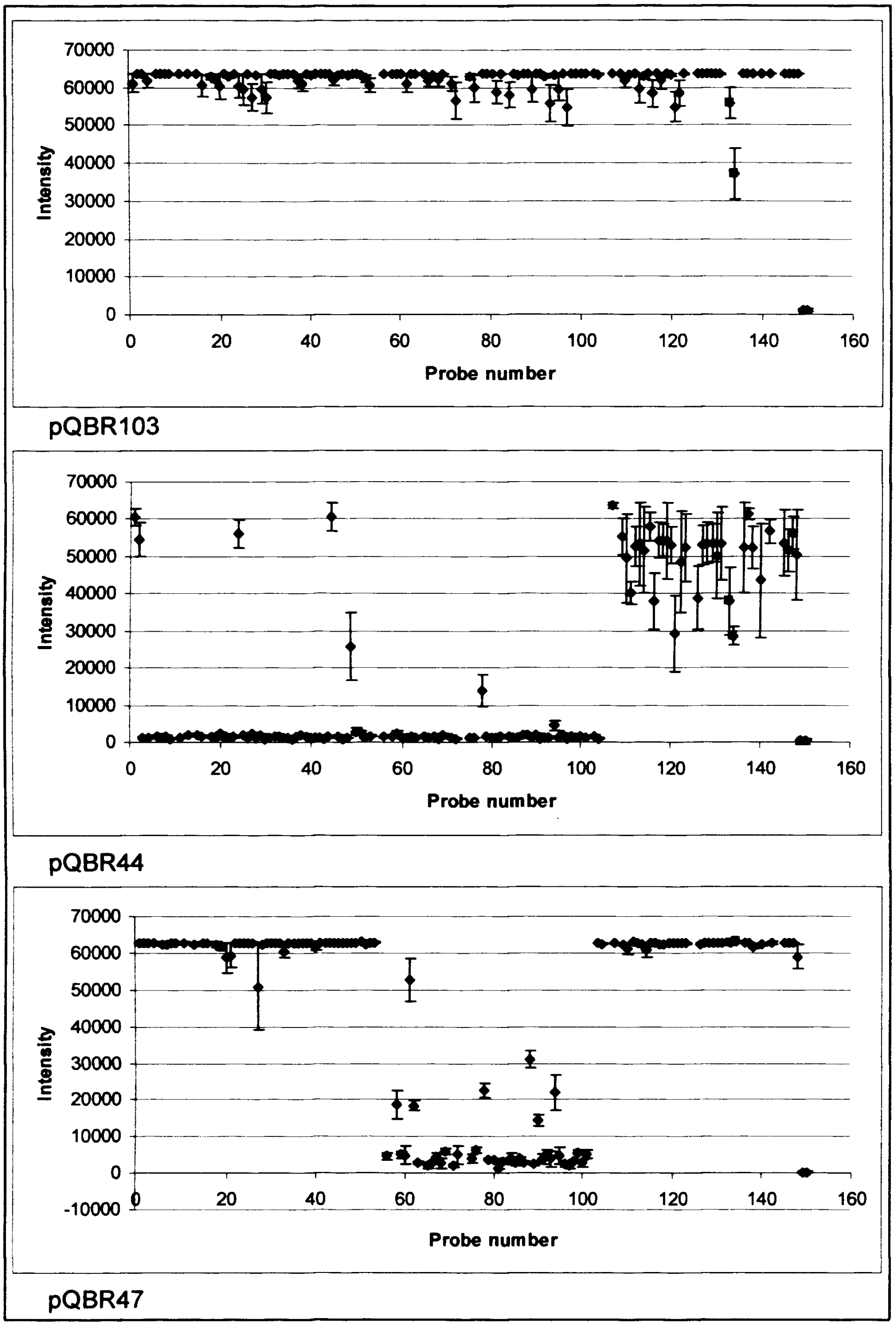
3.4.2.3.2 Identification and characterisation of the pQBR group I common core

The pQBR group I core can be defined as the region that is common to all group I plasmids tested (pQBR103, pQBR44 and pQBR47). This region spaned from ~300, 000 bp to ~5, 500 bp (Figure 3.5) of the pQBR103 genome.

In addition to the group I plasmids tested in this chapter it was possible to extrapolate that this core is present in other group I plasmids (pQBR4 pQBR41 and pQBR42) that have been isolated, based on the PCR surveys undertaken in chapter 2. For all the 35 loci tested in the surveys, pQBR4, pQBR41 and pQBR42 were indistinguishable from pQBR103. Additional sequencing and comparison of shared loci between these plasmids failed to detect any sequence divergence (chapter 2, section 2.4.5). Combining these findings with the estimated genome sizes of pQBR4, pQBR41 and pQBR42 (all are of similar size to pQBR103) (Lilley et al., 1996) it was proposed that these plasmids largely encompassed the pQBR103 genome.

The defined pQBR group I core region represented most (if not all) the pQBR44 genomic sequence (Table 3.3). As such, the pQBR group I core and pQBR44 genome were essentially synonymous. Ergo, maintenance functions and phenotypes known for pQBR44 were attributable to this core region. Figure 3.5 shows the plasmid maintenance and transfer functions annotated for pQBR103 and how they relate to the group I core. While the predicted partition and replication sequences lay within this region, the homology pQBR103 had to the *oriV* of pQBR11 (Viegas et al., 1997) and putative type IV like transfer genes/transfer candidates did not.

A further observation of this "core" region was that it contained a higher density of genes ascribed a putative function compared to the region of pQBR103 not common to the group I plasmids (chi-squared test $p < 0.001$).



pQBR103

pQBR44

pQBR47

Figure 3.4a. Legend given overleaf.

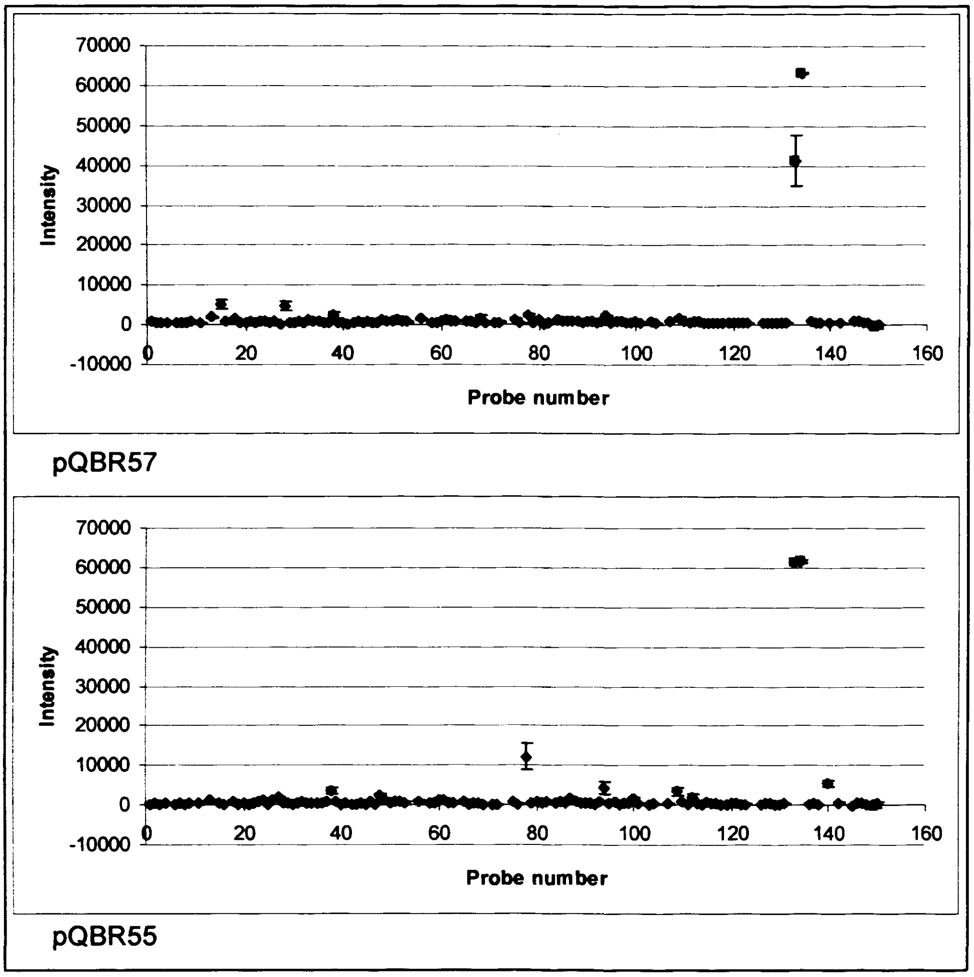


Figure 3.4b

Figure 3.4. Graph of hybridisation intensity of each of the probes on the pQBR103 custom array when challenged with A) pQBR103, pQBR44, pQBR47 B) pQBR55 or pQBR57. Data is only shown for 1 of the two slide replicates performed for each test DNA. Intensity of probe is the average of the median pixel intensity for each replicate spot, standard errors are shown for each probe. Probe number refers to the sequential order probes distributed around the pQRR103 sequence. Diamond data points (Blue) refer to test probes. Square data points (Green) are probes representing the regions of mercury operon (positive controls). Diamond data points (Red) are pUC19 plasmid DNA and *E. coli* DH10 genomic extract (negative controls).

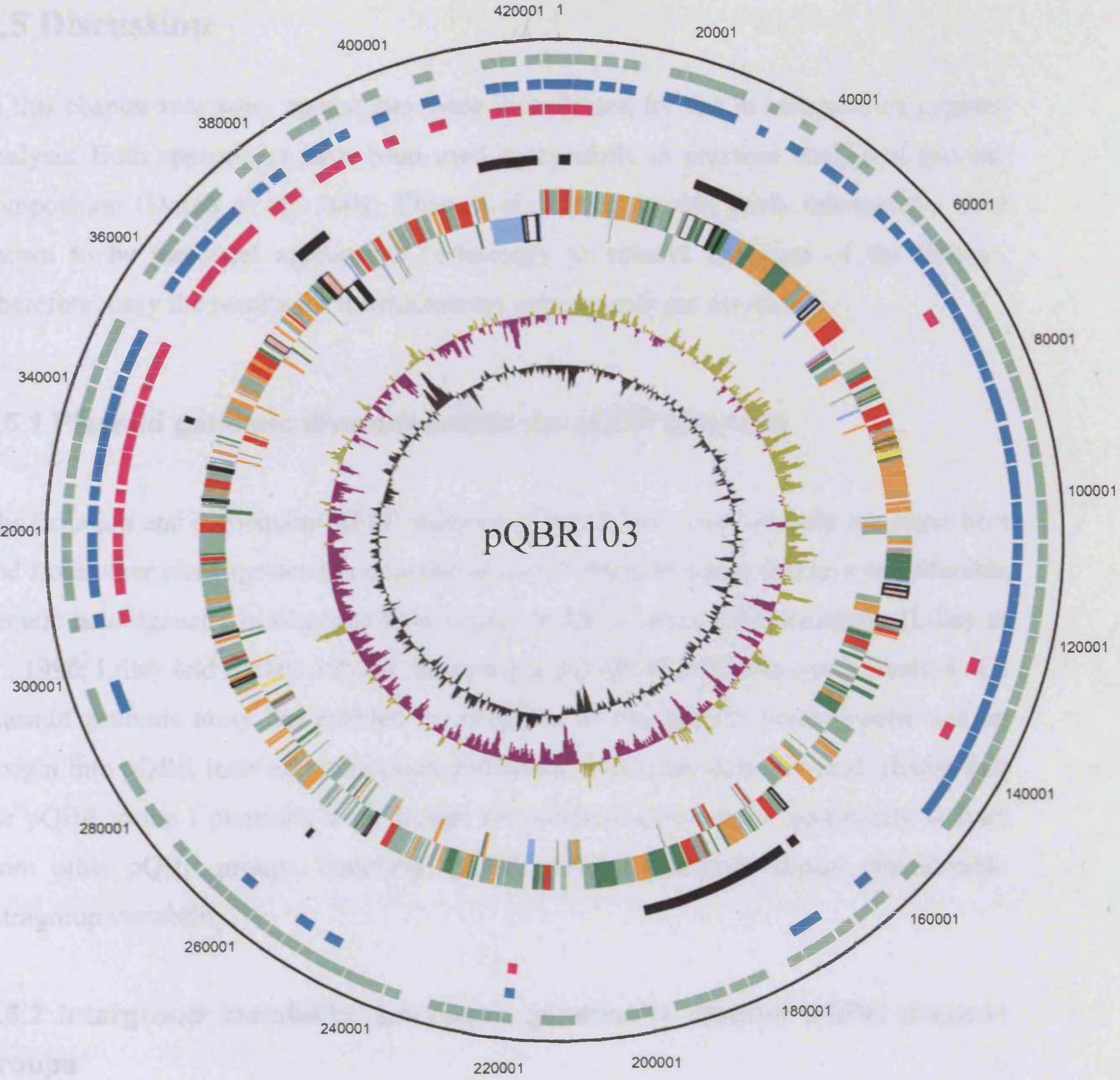


Figure 3.5. pQBR103 microarray results: Regions of commonality detected between group I pQBR plasmids. Outermost circle, **Circle 1** (black) represents the pQBR103 sequence, numbers refer to sequence position (bp). **Circle 2** (green) All 122 probes representing the pQBR103 genomic array and all detected as positive when challenged with pQBR103. **Circle 3** (light blue) probes detected positive when challenged with pQBR47. **Circle 4** (pink) probes detected positive when challenged with pQBR44. **Circle 5** (black blocks) regions on pQBR103 of maintenance function or other interest (clockwise from pQBR103 sequence position 1) 1, plasmid partitioning region; 2, region of transmembrane / secretion homologues; 3, UV resistance genes; 4, region covering Type IV transfer like homologues; 5, region of homology to *oriV* of pQBR11; 6, replication initiation protein; 7, Tn*5042* like transposon and mercury resistance operon; 8, region of chemotaxis like homologues. **Circle 6 and 7** (multicoloured) CDS coded on forward or reverse strand respectively. Colours refer to annotated function of the gene (listed in Figure 2.1 in chapter 2) **Circle 8** (gold and purple) G+C skew, **Circle 9** (black) GC deviation.

3.5 Discussion

In this chapter two array approaches were investigated for use in comparative plasmid analysis. Both approaches have been used successfully in previous studies of genome composition (Dorell et al., 2001; Zhao et al., 2005). In this study microarrays were shown to be the most appropriate technology to achieve the aims of the chapter. Therefore, only the results for the microarray experiments are discussed.

3.5.1 Plasmid genomic diversity within the pQBR plasmids

The isolation and subsequent RFLP analysis of the pQBR plasmids from the sugar beet and from other plant species has demonstrated that this population displays considerable genetic heterogeneity in response to temporal and environmental fluctuations (Lilley et al., 1996; Lilley and Bailey, 1997a). Sequencing pQBR103 and subsequent creation of a plasmid genomic array has enabled investigation of this genetic heterogeneity and an insight into pQBR inter and intragroup dynamics. This study demonstrated, firstly, that the pQBR group I plasmids even though co-localised appear to be genetically distinct from other pQBR groups. Secondly, pQBR group I plasmids display considerable intragroup variability.

3.5.2 Intergroup variability: Localised genetically distinct pQBR plasmid groups

A previous estimation of inter group variability was restricted to the resolution of reciprocal plasmid genome hybridisations (Lilley and Bailey, 1997a). This study suggested that there was little sequence homology between group I and IV pQBR plasmids and hence little intergroup mixing. In this chapter using pQBR103 as a group I representative has corroborated these earlier findings (to a higher resolution) and also extended them to the group III. Two possible explanations can be given for the lack of intergroup mixing: the plasmids are incapable of genetic exchange as they are physically separated by differing host range; and/or these groups offer adaptation to differing niches (niche partitioned), which is reflected in the genomic content.

With respect to host range restrictions this is unlikely, as all pQBR plasmids have been shown to share similar host range by both exogenous and endogenous survey (Bailey et al., 2001) and all can be maintained in pseudomonad hosts. Furthermore, as the pQBR plasmids are recovered from the same habitat, it is reasonable to expect at least some plasmids will transfer between bacterial hosts with similar genetic backgrounds (Bailey et al., 2001). Therefore, even if this co-localisation in a host is rare, there is the opportunity for genetic exchange. The shared mercury transposon between groups I, III and IV may offer some evidence for intergroup mixing. As host range restriction, in itself, is unlikely to describe the observed lack of intergroup mixing, a better explanation is niche partitioning. This suggestion is supported by temporal isolation frequencies of different plasmid groups over a growing season (Lilley et al., 1996; Lilley and Bailey., 1997a), that have shown that different plasmids groups are correlated with the stage of plant maturation. Group I plasmids were found to be temporally separated from both the group III and group IV plasmids. Therefore, the lack of shared sequences observed in this chapter may reflect temporally induced physical separation of group I from other pQBR groups, leading to isolated populations and hence independent evolutionary histories.

3.5.3 Intragroup relatedness: modulation and recombination

In this chapter, little evidence of intergroup exchange was identified. In contrast, considerable intragroup variability was observed between the group I plasmids investigated. In effect the two small pQBR group I plasmids (pQBR44 and pQBR47) are largely a subset of pQBR103, as genome size estimates by RFLP analysis and the plasmid array are congruent, suggesting they have little if any additional DNA to that represented on the microarray. Based on the results of this chapter a simple model can be proposed as to how these genomes relate to one another: pQBR103 gives rise to pQBR47 by a single deletion event, which in turns gives rise to pQBR44 by a further deletion. Of course, the converse is as equally likely i.e. the insertion of DNA into the smaller plasmids to form pQBR103. The identification of structural instability within the pQBR group I, in itself is not unexpected, because of the RFLP analysis of this group. What was surprising by this glimpse into the population genetics of the pQBR group I plasmids was the mechanism of this instability. Analyses of the pQBR103 genome at the regions where the indels were postulated to occur (the breakpoints) identified no specific mechanism for these rearrangements i.e. no IS/transposon or regions of extensive homology (>15 bp). This goes against the conceptual plasmid paradigm of what drives plasmid instability and plasmid genomic structure.

3.5.4 The plasmid paradigm and pQBR group I

Modulistic or mosaic plasmid structure has been studied and discussed for many plasmid groups. Examples include, IncFII (Osborn et al., 2000), IncH (Gilmour et al., 2004) and the *repABC* rhizobial plasmids (Gonzalez et al., 2003). However, probably the best studied group that exemplifies this modular model is the IncP plasmid group (Dennis, 2005). Here functional modules are effectively clustered into regions. Those modules that are of plasmid "housekeeping" function (replication, transfer, partitioning etc) are clustered into the plasmid "backbone". Other modules of transient benefit e.g. catabolic pathways are clustered in adaptive regions (Thomas, 2000). These adaptive regions are highly versatile reflected by the high density of IS and other repetitive elements. Consequently HGT and rearrangements within these regions are frequent (Harayama, 1994; Top and Springael, 2003).

In comparison to the modulistic model, pQBR103 and other pQBR group I plasmids differ in two ways. Firstly, analysis of the pQBR103 genome found it to be conspicuous in the lack of highly repetitive elements and regions of obvious recent HGT. In total only three transposases were identified in the pQBR103 sequence, all of which were associated with the mercury transposon operon (this operon represented the only detectable recent HGT event). For illustration, on average each incP-9 genome completely sequenced has 5.33 open reading frames with homology to transposase sequences (Greated et al., 2002; Dennis and Zylstra, 2004; Sota et al., 2006), even though the average IncP-9 plasmid genome size (93 kb) is considerably smaller than pQBR103's (425 kb).

Secondly, plasmid "housekeeping" functions appear not to be tightly clustered in pQBR103. Although a number could be attributed to a particular region of the genome, this did not constitute a "backbone" as defined by Thomas (2000). Nevertheless, a common core was identified for the pQBR group I plasmids. This common core is proposed to constitute the entire pQBR44 genome, hence the phenotypes associated with this plasmid are attributable to this region. Within this core some annotated pQBR103 plasmid housekeeping functions were identified (replication and partitioning) however, other putative plasmid housekeeping functions were not. For example, the Type IV like transfer genes (chapter 2). As it is known pQBR44 is self conjugative this suggests either an undiscovered transfer system exists within the common core, or that pQBR44 is not just a simple subset of pQBR103, and encodes a divergent transfer system.

Continuing with the IncP-9 as a representative of the modular model of plasmid organisation, the dissimilarity of pQBR103 to this model could be explained by the forces of selection operating on these plasmid genomes. For the IncP-9 plasmids isolated from heavily contaminated environments selection is likely to be high. In comparison, pQBR plasmids were isolated from natural populations of the phytosphere, where forces are likely to be broader and weaker. For the IncP-9 plasmids under high selection, selective sweeps will favour tightly clustered and efficient catabolic pathways as well as stable and efficient backbones, as observed. However, fluidity will also be required to create novel or more efficient pathways and this is provided by IS elements and transposons associated with the adaptive regions. In contrast, pQBR103 and other group I plasmids with broader selection have less pressure for rapid and directed adaptation. Thus, the density of repetitive elements is lower, related functions are less constrained to a particular region and plasmid genome size is larger.

There exists a counter argument. If selection on the pQBR plasmids is generally weak and heterogeneous it could be expected that a tolerance for HGT and rearrangement events exist. Thus, pQBR103 would be expected to have more repetitive elements and evidence of HGT. However, the pQBR103 genome was conspicuous by the absence of such elements and events compared to rhizobial plasmids. Rhizobia are isolated from similar habitats to pQBR plasmids; therefore, it is likely they are under similar selective forces. Rhizobial plasmids do not tightly fit the modular model to the same extent as the IncP. They do, however, show considerable evidence of HGT events and rearrangements mediated by transposons and IS elements, as well as clustering of related functions e.g. nodulation, nitrogen fixation (Giuntini et al., 2005; Gonzalez et al., 2003), in contrast to pQBR103.

3.5.5 Intragroup mixing and plasmid genomic structure remains elusive

Having only a single pQBR plasmid genome sequence has precluded the comprehensive analysis of intragroup mixing that has been described for other plasmid groups. Nonetheless, by comparing the plasmid sequence with the array data it has at least afforded a glimpse into the intragroup dynamics of the pQBR group I plasmids. What is noticeable from this comparison appears to be contradictory. Based on the pQBR103 genomic sequence alone this group appears to be recalcitrant to HGT and rearrangement. In contrast, the array clearly demonstrated this group to be reassorting. Furthermore, this level of reassortment may be greatly underappreciated. At the beginning of this discussion a simple model of group I evolution was given. This model fails to account for some of the probes that hybridised for pQBR44 and pQBR47 that lay in the regions postulated to be absent from these plasmids, referred to as "anomalous" probe hybridisations. Two explanations can be given for these positive probes. Firstly, this was probe cross hybridisation, or secondly, the model given was an oversimplification and considerably more rearrangements have happened within the group I plasmids.

With respect to cross hybridisation this was a limitation of amplifying inserts from a sequencing clone library by universal primers. Although this provides the most cost effective method for constructing probes (compared to oligonucleotide synthesis or targeted PCR amplification approaches), probe size was dictated by insert size. For pQBR103 the average probe size was 2400 bp. This probe size was considerably larger than other PCR CGH arrays where probes were specifically amplified (size 200-1000 bp) (Salma et al., 2000; Fitzgerald et al., 2001; Smoot et al., 2002). Obviously, as probe size increases so does the likelihood of probe cross hybridisation. However, in this study the "anomalous" probe hybridisations did not show cross hybridisation to host, and based on *in silico* analysis, to other pQBR103 regions (with the exception of one, see results). Therefore, it is likely that these hybridisation signals were attributable to the plasmid sequences.

The previous PCR study and sequencing of shared loci (chapter 2) indicated that the comparison of pQBR plasmids would likely give rise to qualitative data, hence, a single channel hybridisation method was used (this also greatly reduced the cost of the technology). However, a limitation to this design was that the level of divergence between target and probe cannot be as accurately determined, as is afforded by a two channel system (control and test hybridised simultaneously). Therefore, some of these "anomalous" hybridisations may in fact be false positives or highly divergent sequences (those with low hybridisation intensity) that could not be resolved using the one channel array. However, others (with high hybridisation intensity) can confidently be given as evidence of a loss of synteny between the group I plasmid sequences i.e. indicative of a potential recent internal rearrangement. Therefore, it is likely the model presented earlier was an oversimplification; Figure 3.6 takes into consideration internal rearrangements and acquisition of divergent sequences.

Using the pQBR103 custom array the level of synteny can only be assumed, not proven i.e. it is unknown whether the group I core is syntenic to all members. Therefore, the true extent of rearrangements cannot be determined using an array. Regardless of the extent, what is clear form this chapter is that reassortment does occur in the group I and in the most part this is unlike that described for other plasmids i.e. not largely mediated by MGEs. This is intriguing as it suggests that instead of having directed and frequent mixing this is likely to be infrequent and largely a stochastic process in the pQBR group I plasmids. However, it must be reiterated that this study has only offered a glimpse into the population genetics of the group plasmids; as such the significance of these findings remains elusive.

3.6 Future work

The major drawback to all comparative genomic studies using arrays is that it is not possible to detect novel sequences. In this study it was found that plasmid representatives from group III and group IV had little similarity to pQBR103. Although this finding was informative in assessing intergroup relatedness, effectively this meant the array was fruitless in furthering the understanding of these groups. Therefore, the next logical step would be to sequence representatives from these groups. Analysis of these plasmid sequences would not only give a better understating of their biology individually, but in comparison provides a further insight into intergroup dynamics and collective phytosphere fitness (discussed further in chapter 6). In addition to sequencing representatives from other pQBR groups, the sequencing of multiple members of a group is also likely to be informative. For example, sequencing of pQBR44 and pQBR47 would establish the true level of intergroup dynamism as well as provide a better understanding of the mechanisms governing the process. Also, by detecting what is common and novel to each plasmid this would enable an estimation of the genetic and functional diversity within this group.

Sequencing is the obvious way forward. However, if this option were not possible further insights into the mechanism of intragroup mixing could be gained by identifying the "breakpoints" on pQBR103 where the "indel" events were predicted to take place. This might be achieved by primer walking and sequencing out from the boundaries of homology pQBR44 and pQBR47 share with pQBR103.

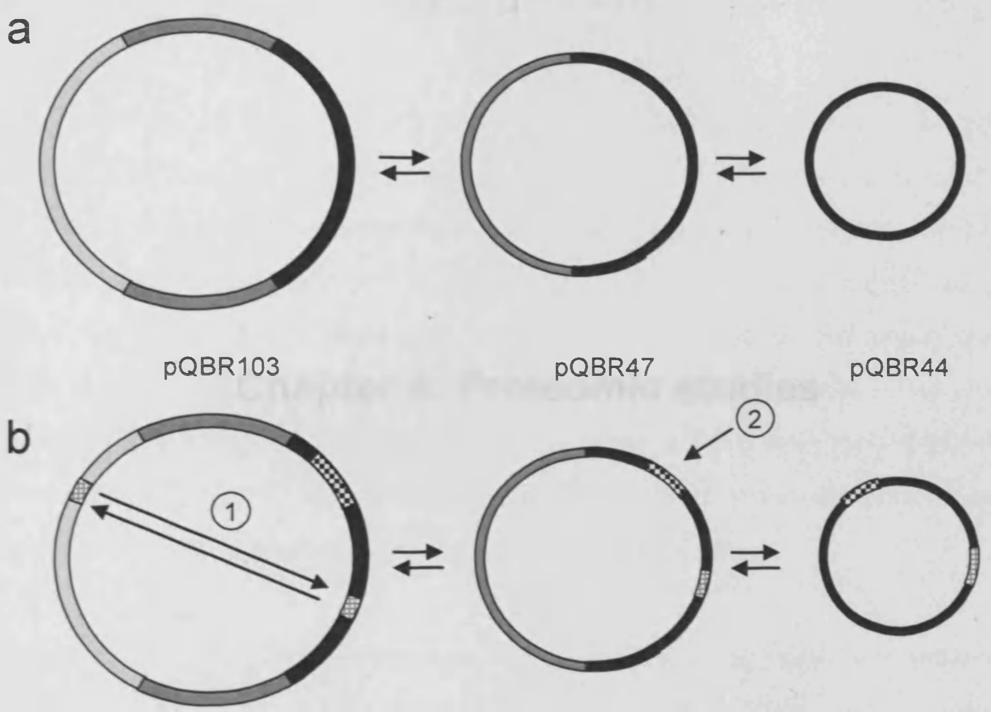


Figure 3.6 Group I rearrangements: Hypotheses of how the group I plasmids pQBR103, pQBR47 and pQBR44 relate to one another. a) a simplified model b) Model including potential internal reassortments and HGT events. For both Figures, **black** represents sequence common to all three plasmids; **dark grey** regions are common to pQBR47 and pQBR103; **light grey** shows sequence unique to pQBR103. Plasmid sizes and regions of shared sequences are illustrative only and not to scale. **Fig a**. At the simplest pQBR47 and pQBR44 can be seen as subsets of pQBR103, either of which can arise by single deletion event or sequential deletions (as shown). Alternatively the converse is also likely (insertions into the smaller plasmids can give rise to pQBR103). However, this model does not account for positive probe hybridisations in the regions postulated to be absent from pQBR44 and pQBR47 (shown of Figures 3.4 and 3.5). **Fig b**. Possible reasons for these “anomalous” hybridisations are: Firstly, these are false positives caused by cross hybridisation, Secondly, they are hybridising to sequences present in pQBR44 and/or pQBR47 that are evolutionary divergent from pQBR103 and have been obtained by an HGT event (number 2 on Figure, and shown by **chequered boxes**). Thirdly, internal rearrangements occur which give rise to loss of synteny between the group I plasmids (number 1 on Figure and indicted by **hatched boxes**). This model offers an explanation for the findings shown in Figure 3,4 and 3.5. However, the extent of internal rearrangements and potential HGT events into pQBR44 and pQBR47 is undetermined.

Chapter 4: Proteomic studies

4.1 Introduction

In this thesis the investigation of pQBR103 has, so far, been limited to the genotypic level i.e. analysing the genetic capacity of the plasmid through the annotation and analysis of the sequence (Chapter 2). Although such endeavours are sometimes fruitful in enabling discussion of what might contribute to the fitness benefit of plasmid carriage, for pQBR103 this discussion is limited. This limitation was largely due to the amount of DNA that could not be ascribed functions i.e. 80% of CDS are either orphan or conserved hypotheticals. Of this 80% a proportion have been demonstrated to be expressed in response to the environment by previous *in vivo* expression technologies (IVET) (Zhang et al., 2004a; b) and were discussed in chapter 2.

Although IVET technologies have been useful in identifying regions of putative gene expression in the natural environment (Rediers et al., 2005), there are limitations to the technology. Firstly, the technology only qualitatively detects what is induced in the environment. Therefore, regions of the genome that are up or down regulated are not detected. This means information regarding how plasmid and host interact is lost. Secondly, it only identifies regions that are potentially expressed at the level of transcription and not translation. As post transcriptional regulation and processing is not accounted for, the specific proteins that are expressed in that environment can only be inferred. This latter problem is not restricted to IVET but a limitation of all transcriptional studies, including widely used transcriptomic arrays (Guerreiro et al., 1999). As a consequence, it is a worthwhile exercise to detect proteins themselves when investigating the interaction of an organism with its environment.

Proteomic investigations are essentially concerned with separating a complex mix of proteins; usually the protein complement expressed by a genome (the proteome) (Wilkins et al., 1996), combined with the subsequent identification of the proteins, as introduced in chapter 1. Proteomic analysis has been used previously to investigate bacterial plant interactions (Langlois et al., 2003). Proteomic studies investigating the expression of both host as well as accessory plasmid proteins when subjected to plant metabolites have also been reported. Two examples are, the response of *R. leguminosarium bv trifolii* carrying indigenous plasmids to flavonoids (Guerreiro et al., 1997), and recently the response of *A. tumefaciens* carrying the pTi plasmid to acetosyringone, a product released from wounded plant cells (Lai et al., 2006). More specifically to this chapter the response of bacterial strains carrying plasmids and their cured derivatives have been investigated with the aim of identifying environmentally expressed plasmid proteins (Guerreiro et al., 1998). The proteomic effect of pQBR103's carriage has been studied previously (in house, unpublished data). In this study the proteome of *P. fluorescens* SBW25 with and without pQBR103 was compared on the rich solid laboratory media, LBA. Solid media rather than liquid media was used as *P. fluorescens* is a surface coloniser. In this experiment only qualitative differences were investigated i.e. proteins present with pQBR103 carriage and absent without. In total only three differences were observed. These observations suggest that pQBR103 is unlikely to offer any fitness advantage to its host in such nutrient rich conditions, most likely as the conditions are atypical to those experienced in nature. Also, the regulation of pQBR103 is tightly controlled which might be expected to minimize the burden imparted by the plasmid on its host. It was decided to extend these preliminary investigations to different media, R2A, (Reasoner and Geldreich, 1985) and Water agar amended with pea seed exudate, as these are predicted to be more representative to the natural conditions *P. fluorescens* SBW25 +/- plasmid may experience.

4.2 Chapter aims

1) To detect qualitative protein changes in response to differing environments that are (postulated to be) plasmid encoded i.e. present in SBW25/pQBR103 absent from SBW25 only condition by Two Dimensional Electrophoresis (O'Farrell, 1975), and to identify these proteins by Mass spec analysis.

2) To detect quantitative and qualitative expression changes of chromosomally encoded genes as a consequence of pQBR103 carriage in different environments (by Two Dimensional Electrophoresis). Also, identify the proteins that are up-regulated to the greatest extent by pQBR103 using Mass spec analysis.

This will enable the following hypothesis/questions to be addressed.

1) In chapter 2 the large proportion of orphan CDS (60%) were hypothesised to represent genes of unknown ecological function. If so can responses to the simulated environmental conditions (R2A and WA pea seed) be detected that involve orphan proteins, further supporting this hypothesis and the pQBR103 CDS predictions?

2) It is hypothesised that the expression response of pQBR103 and host will be greater on R2A and WA pea seed media than has been previously demonstrated on a rich laboratory medium (LBA). As these media represent closer approximations of the conditions host and plasmid would experience in nature. In particular it is hypothesised that WA pea seed agar will elicit the greatest plasmid/host response as this represents the conditions mostly closely akin to the natural environment pQBR103 experiences. The second question posed is; what is the functional basis of this response, is it a coordinated (global) response between plasmid and host replicon, as predicted by the identification of putative global regulators on pQBR103 in chapter 2?

4.3 Materials and Methods

All reagents were from Sigma-Aldrich, UK or from Oxoid, UK unless otherwise stated.

4.3.1 Media, culturing conditions and cell preparation

4.3.1.1 Preparation of pea exudates and media

Pea seed exudate was chosen for this experiment as it has been shown that *P. fluorescens* SBW25 is a good coloniser of young pea seedlings (*Pisum sativum var. quincy*) (Houlden, 2005). Pea exudate was made as previously described (Timms-Wilson, 1998). To obtain approximately 40 ml of pea exudates, 250 ml of milliQ H_2O was added to 200 grams of pea seed and shaken for 16 hours at 28 °C, before being sieved and filtered using a 0.2 μm filter (Nalgene, Fisher Scientific, UK).

R2A media was made and sterilized as per manufacturer's instructions with and without $HgCl_2$ (7 μg ml^{-1}) selection. Water agar (WA) pea exudate plates, with and without 25% strength $HgCl_2$ selection, were made by adding 20 ml of pea seed exudate to 50 °C autoclave sterilised WA agar that had been made up to 96% volume (480 ml) with milliQ H_2O (correcting for addition of exudates), and used the same day.

4.3.1.2 Culturing and cell preparation for 2-D PAGE

A schematic of culturing conditions and cell preparation is given in Figure 4.1.

Strains *P. fluorescens* SBW25 and *P. fluorescens* SBW25 carrying pQBR103 were recovered from glycerol stocks by plating onto PSA and PSA containing 14 μg ml^{-1} $HgCl_2$ plates, respectively, and incubated at 28 °C for 48 hours. A single colony was streaked on PSA and PSA 27 μg ml^{-1} $HgCl_2$ plates and incubated at 28 °C for 48 hours. Three 5ml LB cultures with 27 μg ml^{-1} $HgCl_2$ where appropriate, were seeded from individual colonies for each of the plasmid-free and plasmid carrying strains, representing three biological replicates. Cultures were shaken at 180 rpm at 28 °C for 16 hours.

For each starter culture 1.5 ml was harvested by centrifugation at 1,400 x *g* at 4 °C for 10 minutes. The supernatant was discarded and the pellet was washed by resuspending in 1ml 4 °C PBS followed by a further 10 minute, 1,400 x *g* centrifugation. The washing step was repeated again before resuspending in 1 ml PBS. Each cell suspension was diluted to an optical density of 1.0 at 600 nm then diluted ten to one thousand fold. One hundred microlitres of each cell suspension was spread on both R2A and WA 4% pea exudates with or without mercury selection, where appropriate. Plates were incubated at 28 °C for 48 hours.

The inoculum applied to the experimental plates was ascertained by making serial dilutions in PBS and plating onto LBA media in duplicate for each biological replicate. Plates were incubated at 28 °C for 24 hours and the resultant colonies were counted.

For each biological replicate a plate was selected that had a bacterial lawn at the highest dilution. i.e. 12 plates in total (3 plates of SBW25 on R2A, 3 plates SBW25 on WA pea exudates, 3 plate of SBW25/pQBR103 on R2A and 3 plates SBW25/pQBR103 on WA pea exudates medium). Lawns were washed from plates by adding 2 ml 4 °C PBS and disrupting the cells with a plate spreader, before removing the suspension with a wide bore pipette. Each suspension was aliquoted into two 1.5 ml microtubes and centrifuged at 1,400 x *g* for 10 minutes at 4 °C and the supernatant removed. Cells were stored by freezing at -80 °C. In total this gave 24 experimental samples.

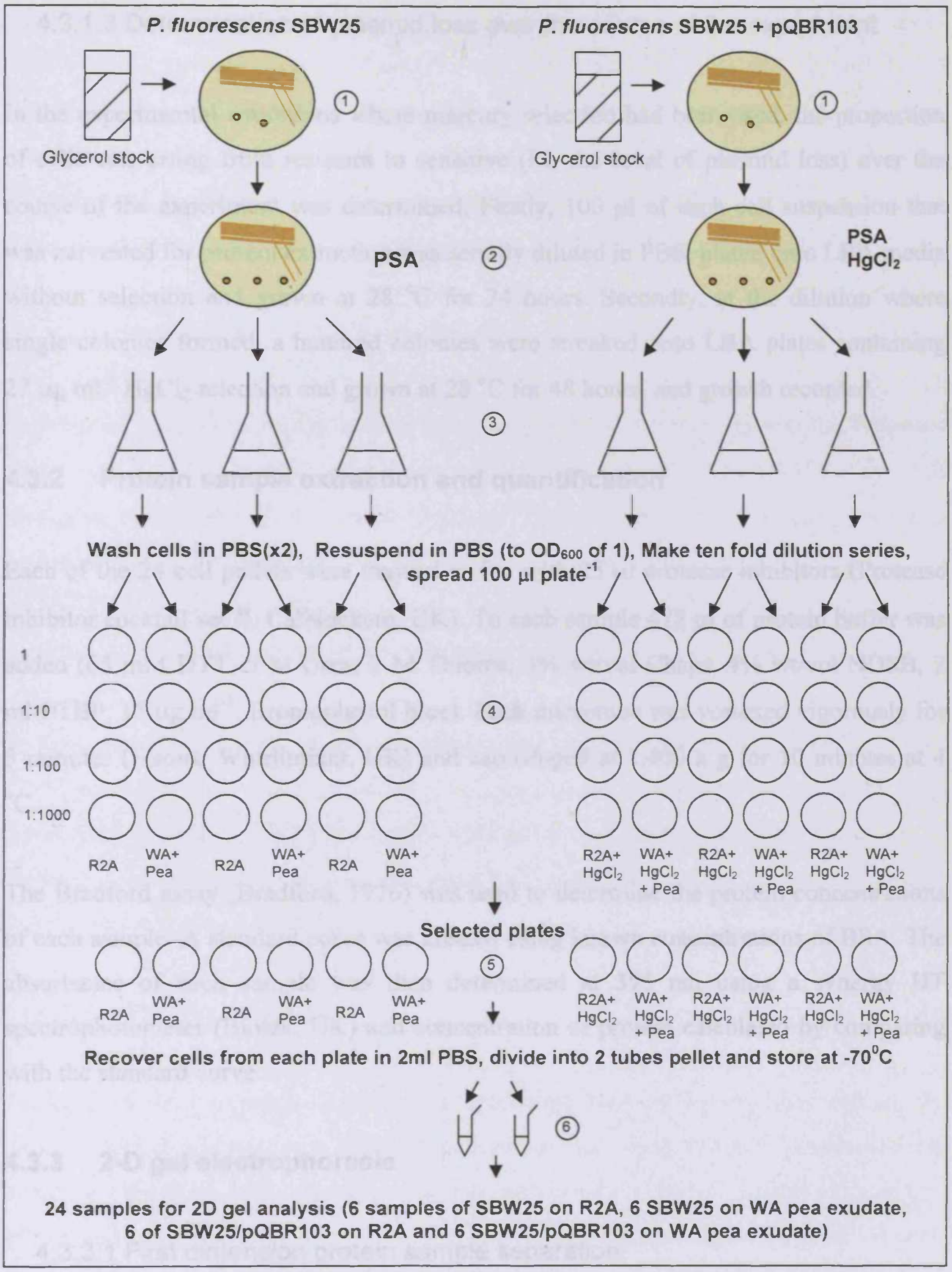

P. fluorescens SBW25
P. fluorescens SBW25 + pQBR103
Glycerol stock
Glycerol stock
1
1
PSA
2
PSA
HgCl2
3
Wash cells in PBS(x2), Resuspend in PBS (to OD600 of 1), Make ten fold dilution series, spread 100 µl plate-1
1
1:10
1:100
1:1000
4
R2A
WA+ Pea
R2A
WA+ Pea
R2A
WA+ Pea
R2A+ HgCl2
WA+ HgCl2 + Pea
R2A+ HgCl2
WA+ HgCl2 + Pea
R2A+ HgCl2
WA+ HgCl2 + Pea
Selected plates
5
R2A
WA+ Pea
R2A
WA+ Pea
R2A
WA+ Pea
R2A+ HgCl2
WA+ HgCl2 + Pea
R2A+ HgCl2
WA+ HgCl2 + Pea
R2A+ HgCl2
WA+ HgCl2 + Pea
Recover cells from each plate in 2ml PBS, divide into 2 tubes pellet and store at -70°C
6
24 samples for 2D gel analysis (6 samples of SBW25 on R2A, 6 SBW25 on WA pea exudate, 6 of SBW25/pQBR103 on R2A and 6 SBW25/pQBR103 on WA pea exudate)

Figure 4.1 Schematic of proteomic experimental design. 1) Recovery from glycerol stocks 2) Plating to single colonies 3) Independent colonies seeded into liquid culture (representing three biological replicates) 4) Each culture was washed, diluted to a uniform optical density, serially diluted and spread onto experimental media 5) Plates at chosen growth density are selected and cells recovered 6) Recovered cells are aliquoted twice to represent experimental replicates

4.3.1.3 Determination of plasmid loss over the course of the experiment

In the experimental conditions where mercury selection had been used, the proportion of cells converting from resistant to sensitive (i.e. the level of plasmid loss) over the course of the experiment was determined. Firstly, 100 µl of each cell suspension that was harvested for protein extraction was serially diluted in PBS, plated onto LBA media without selection and grown at 28 °C for 24 hours. Secondly, at the dilution where single colonies formed, a hundred colonies were streaked onto LBA plates containing 27 µg ml^{-1} $HgCl_2$ selection and grown at 28 °C for 48 hours, and growth recorded.

4.3.2 Protein sample extraction and quantification

Each of the 24 cell pellets were thawed on ice with 25 µl protease inhibitors (Protease inhibitor cocktail set II, Calbiochem, UK). To each sample 475 µl of protein buffer was added (65 mM DTT, 5 M Urea, 2 M Thiorea, 4% wt/vol Chaps, 4% wt/vol NDSB, 2 mM TBP, 15 µg ml^{-1}, Bromophenol blue). Each microtube was vortexed vigorously for 5 minutes (Fisons, Whirlimixer, UK) and centrifuged at 1,400 x *g* for 10 minutes at 4 °C.

The Bradford assay (Bradford, 1976) was used to determine the protein concentrations of each sample. A standard curve was created using known concentrations of BSA. The absorbance of each sample was then determined at 595 nm using a synergy HT spectrophotometer (Biotek, UK) and concentration of protein calculated by comparing with the standard curve.

4.3.3 2-D gel electrophoresis

4.3.3.1 First dimension protein sample separation

Each of the 24 samples were diluted to 100 µg protein in 180 µl protein buffer. Each sample was applied to a 11cm pH 4-7 isoelectric focussing strip (Bio-Rad, UK) by passive rehydration for 16 hours under mineral oil (Bio-Rad, UK). Strips were removed from the oil and equilibrated in buffer I (6 M UREA, 0.375 M Tris-HCL pH 8.8, 2%

SDS, 20% glycerol, 2% DTT) for 10 minutes. The buffer was removed and replaced with buffer II (6 M UREA, 0.375 M Tris-HCL pH 8.8, 2% SDS, 20% glycerol, 2% iodoacetamide) and left to equilibrate for another 10 minutes. Proteins on the strips were focussed under mineral oil on a Bio-Rad Protean IEF cell using the following conditions: at 20 °C, step 1, linear ramp to 250 V for 20 minutes; step 2, linear ramp to 8,000 V for 150 minutes; step 3, rapid ramp to 8,000 V for a total of 25,000 V-hr.

4.3.3.2 Second dimension protein sample separation

Separation of proteins in the second dimension was by gel electrophoresis. Focussed IEF strips were embedded into Criterion™ XT precast 12% Bis-Tris 11 cm gels alongside 1 µl protein marker standard (Mark 12, Invitrogen, UK) in 1x MOPs running buffer (Bio-Rad, UK) at 200 V for ~1.5 hours.

4.3.3.3 Protein fixing, staining and image capture

Protein gels were fixed, stained and washed by gentle agitation, firstly in fix buffer (10% methanol, 7% acetic acid) for 30 minutes before replacing fix buffer with SyproRuby (Bio-Rad, UK) for 3 hours, and finally washing the gel in fix buffer for 30 minutes. Gel images were captured using a Bio-Rad FX imager fluorescent protein indetection at medium sensitivity.

4.3.3.4 Protein gel analysis

For each of the two media conditions investigated, R2A or WA and pea extract, the process of gel analysis was identical and performed using PDQuest basic excision software (Bio-Rad, UK). Firstly, variation between gels due to image capture was eliminated by normalisation function in the software package. Secondly, proteins were matched between the six SBW25 only and six SBW25 carrying pQBR103 gels by a combination of automatic and manual matching. Proteins present in *P. fluorescens* SBW25 gels and absent from all *P. fluorescens* SBW25/pQBR103 gels, or vice versa were recorded as qualitative changes. Because of the number of replicate gels, proteins that showed increase or decreased expression of at least two fold in response to pQBR103 carriage were also recorded.

4.3.4 Protein identification

4.3.4.1 Protein excision and digestion

Spots chosen for excision were excised from the gels using the Proteome Works™ Spot Cutter System (Bio-Rad, UK). Replicate spots from each gel condition were pooled in a single well of a 96 well plate. Proteins spots were pooled in a strategy to increase the chances of protein identification due to weak protein concentrations.

Each collection of excised spots was washed twice in 50 μl ammonium bicarbonate (25 mM) before being dehydrated by the addition of 5 μl acetonitrile (HPLC grade) for 15 minutes. The acetonitrile was removed and samples were rehydrated by adding 50 μl ammonium bicarbonate (25 mM) for 10 minutes and then the excess ammonium bicarbonate was removed. Five microlitres acetonitrile was added for 15 minutes in a final dehydration step. Samples were digested in 30 μl diluted sequencing grade modified trypsin (Promega, UK) for 37 °C overnight. The amount of trypsin varied depending on protein concentration (0.33 ng – 6.67 ng μl^{-1}).

The peptides were removed from the gel plugs by adding 30 μl extraction solution (53 μl formic acid 100 μl acetonitrile made up to 5 ml with HPLC grade water) and leaving for 30 minutes at room temperature. The solution for each sample was transferred to a 0.5 ml Eppendorf tube and concentrated to ~1 μl using an Eppendorf 5301 concentrator (Eppendorf, Germany). Samples were then ready for mass spectrometry analysis.

4.3.4.2 Mass spectrometry analysis (MALDI - TOF/TOF)

Samples were run by Dr Xiaohong Yu, a mass spectrometry specialist at the Centre for Ecology and Hydrology, Oxford.

Tryptic peptides were dissolved in 5 µl 0.1% TFA (trifluoroacetic acid), and then desalted using ZipTip C18 (Millipore, Bedford, MA, USA). Peptides were eluted directly onto a 192-well MALDI plate using an elution solution consisting of 1:3 CHCA matrix solution (10 mg ml^{-1} α-cyano-4-hydroxycinnamic acid (CHCA) in 50/50 acetonitrile/0.1 TFA) and 80/20 acetonitrile/0.1TFA. All sample analysis was carried out on a 4700 Proteomics Analyzer, MALDI-TOF-TOF- mass spectrometer (Applied Biosystems, Foster City, CA, USA) using 4000 Series Explorer Software v3.0. The instrument was calibrated using the CalMix (ABI 4700 Proteomics Analyzer Calibration Mixture). MS data were automatically acquired with reflector-positive mode to generate a peptide mass fingerprint (PMF). The six most intense ions were selected using an exclusion list of trypsin autodigestion and keratin peptides for automated MS/MS analysis. Nitrogen air was used as the collision gas and the collision energy was set at 1 kV.

4.3.4.3 Database searches

Data from MALDI MS and MS/MS acquisitions were used in a combined search against the pQBR103 protein database using MASCOT (Matrix Science, London,UK) with the following setting: all entries, peptide tolerance at 1.0 Da, MS/MS tolerance at 0.8 Da, carbamidomethylation of cysteine (fixed modification), and methionine oxidation (variable modification).

4.4 Results

4.4.1 Two dimensional protein gel analysis

4.4.1.1 Protein detection and gel reproducibility

The protein expression profile of *P. fluorescens* with and without pQBR103 was compared in two separate experiments. Firstly, a comparison was made in response to being grown on the low nutrient medium, R2A. Secondly, a comparison when grown on WA supplemented with 4% pea seed extract. In both experiments, where inorganic mercury selection was used to maintain plasmid in host, no plasmid loss was detected.

In Table 4.1 the level of protein detection and reproducibility within and between condition is given for experiment 1, growth on R2A medium and experiment 2, growth on WA + pea seed.

a) R2A medium

Condition	Number of gels	Number of proteins detected	Number of proteins common to all gels	Mean Coefficient of variance
SBW25 + pQBR103	6	404	194	31.68
SBW25 Only	6	397	234	29.9
Overall	12	407	139	30.84

b) WA + 4% Pea seed exudate medium

Condition	Number of gels	Number of proteins detected	Number of proteins common to all gels	Mean Coefficient of variance
SBW25 + pQBR103	6	450	275	33.34
SBW25 Only	6	450	305	32.87
Overall	12	453	236	33.34

Table 4.1. Protein detection variability observed within and between conditions, SBW25 +/- pQBR103 when grown on a) R2A medium b) WA + pea seed exudate

4.4.1.2 Detection of proteins only expressed in the presence of pQBR103

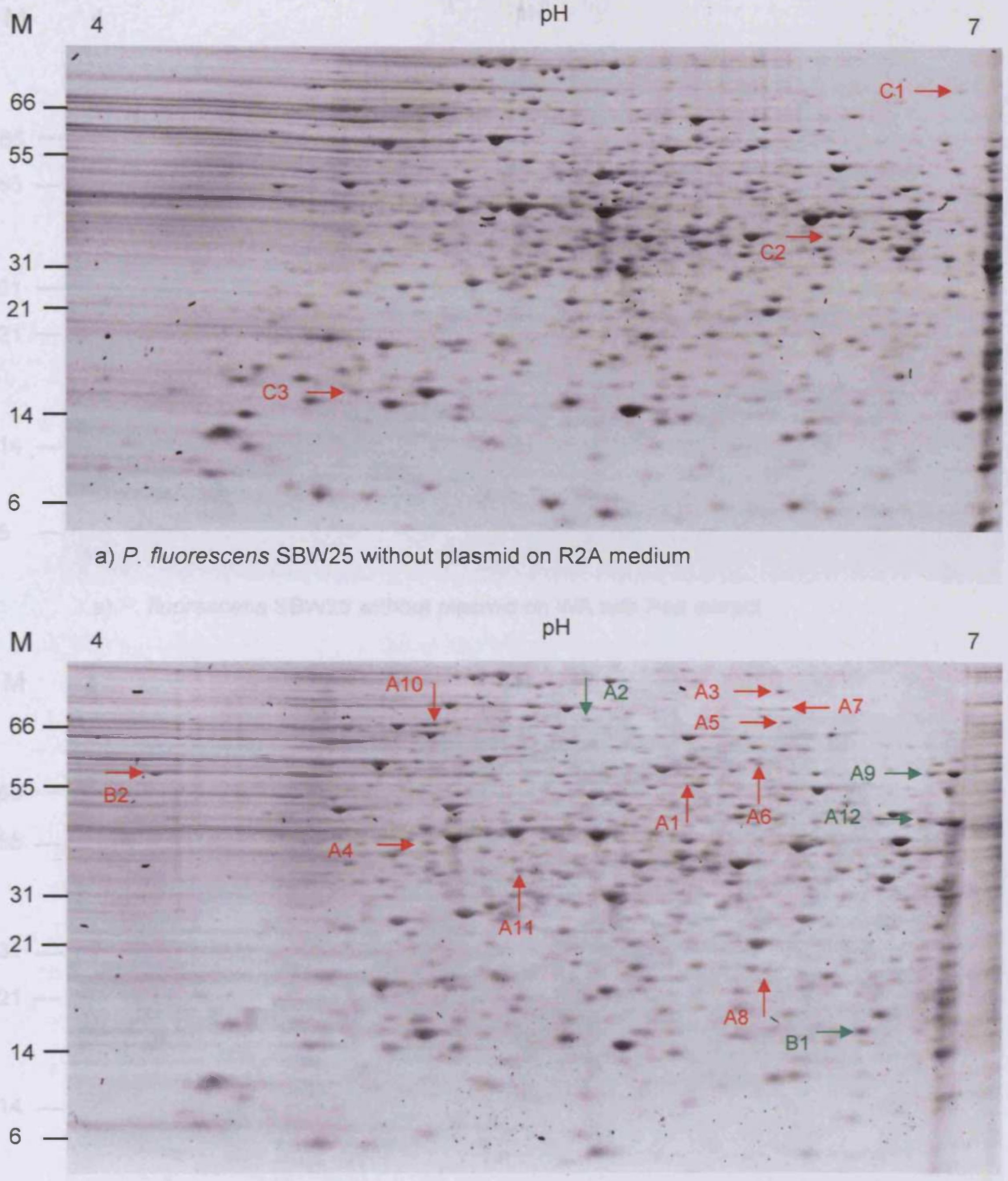
In both experiments the total number of proteins expressed that were likely to be encoded by pQBR103 was 13. (Table 4.2). By far the largest number of these proteins were detected when *P. fluorescens* SBW25/ pQBR103 was grown on a R2A medium. Comparisons of the proteins detected between the two conditions were made. Based on isoelectric point and protein size the three proteins detected in the WA pea seed environment were common to the R2A condition (shown in Figures 4.2 and 4.3). For one of these proteins this was confirmed by Mass spectrometry (see section 4.4.2.2). In total, therefore, a maximum of 10 potential pQBR103 encoded plasmids were detected by 2-D analysis.

	Experimental medium	
	R2A	WA pea exudates
Present in SBW25/pQBR103 Absent in SBW25	10	3
Average number of replicate gels each protein was detected in (max 6)	6	5.7

Table 4.2. The number of unique proteins detected when *P. fluorescens* SBW25 was carrying pQBR103 in comparison to the plasmid free strain. Data is given for growth on two separate media, R2A or WA pea seed supplemented medium. The average number of replicate gels each protein was identified in (max 6) is given for the experimental media.

4.4.1.3 Detection of changes in *P. fluorescens* SBW25 protein expression as a result of pQBR103 carriage.

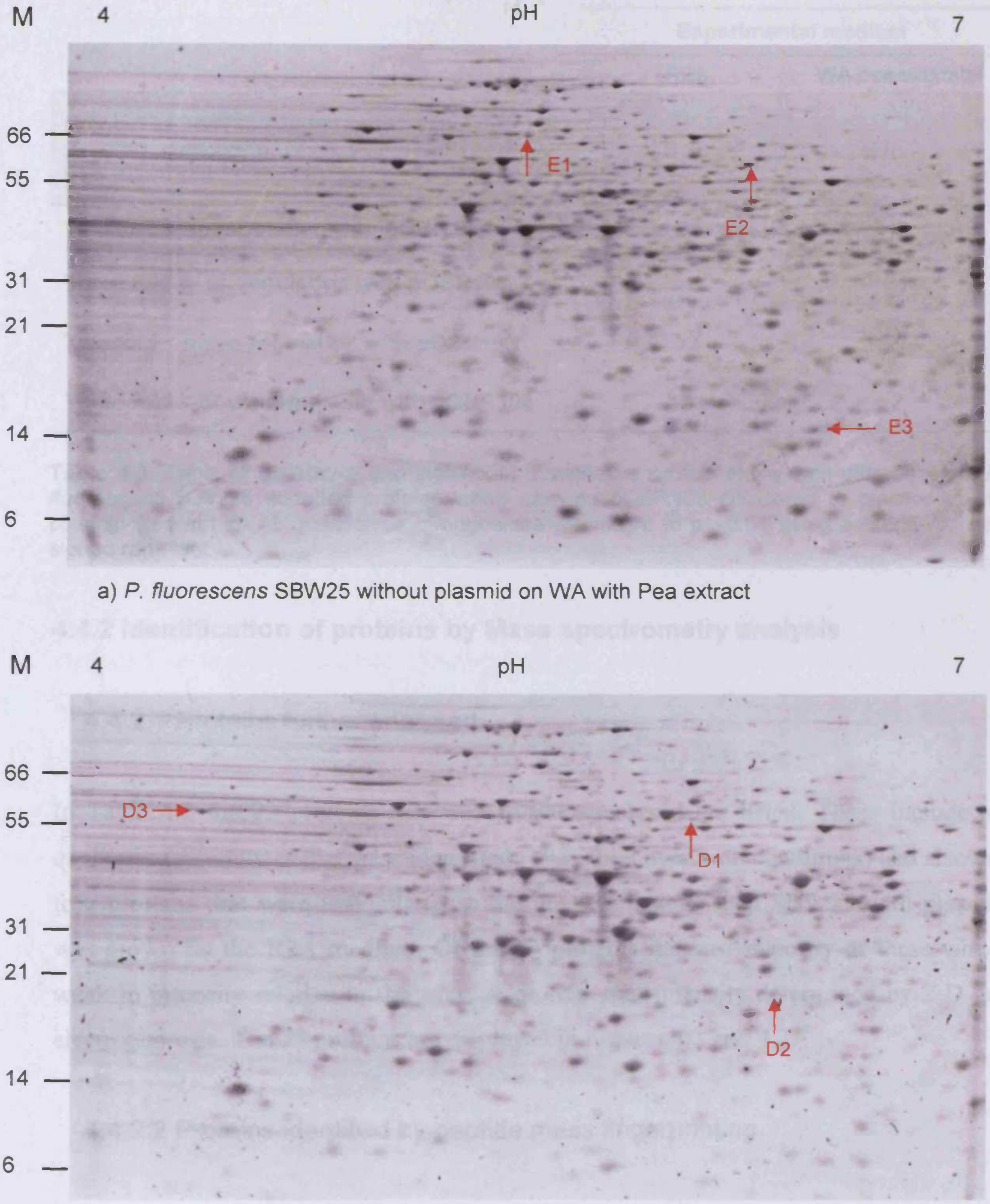
In Table 4.3 the number of chromosomally encoded qualitative and quantitative changes associated with pQBR103 carriage is given for different experimental environments. In both environmental conditions three qualitative changes in host protein expression were identified. Comparison of the isoelectric point and protein size identified these changes to be unique to each of the environmental conditions (Figures 4.2a and 4.3a), suggesting these changes were a response of plasmid and host to differing environments, not a host response to plasmid carriage. In addition to qualitative changes a significant alteration in the expression levels of host encoded proteins was identified with plasmid carriage in both experimental conditions (Mann-Whitney signed rank test $p < 0.01$).



a) *P. fluorescens* SBW25 without plasmid on R2A medium

b) *P. fluorescens* SBW25 carrying pQBR103 only on R2A medium

Figure 4.2. Two-Dimensional protein gel from proteins extracted from *P. fluorescens* SBW25: grown on a R2A medium a) without pQBR103 and b) with pQBR103. Red arrows indicate all qualitative protein differences and green arrows the largest quantitative differences caused by plasmid carriage. Numbers refer to spot identifiers used in Table 4.4 M indicates size markers, sizes are shown in KDa. Note, only one gel of six replicates per condition is shown.



a) *P. fluorescens* SBW25 without plasmid on WA with Pea extract

b) *P. fluorescens* SBW25 with pQBR103 on WA with Pea extract

Figure 4.3. Two-Dimensional protein gels from proteins extracted from *P. fluorescens* SBW25: grown on WA supplemented with pea extract a) without pQBR103 and b) with pQBR103. Red arrows indicate all qualitative protein differences detected. Numbers refer to spot identifiers used in Table 4.4 M indicates size markers, sizes are shown in KDa. Note, only one gel of six replicates per condition is shown.

	Experimental medium	
	R2A	WA pea exudate
Present in SBW25 Absent in SBW25/pQBR103 (Average number of replica gels each protein was identified in, max 6)	3 (6)	3 (5.7)
Two fold < up regulation with pQBR103	22	10
Three fold < up regulation with pQBR103	4	11
Two fold < down regulation with pQBR103	23	27
Three fold < down regulation with pQBR103	5	2

Table 4.3. Table of qualitative and significant quantitative protein expression differences in *P. fluorescens* SBW25 encoded proteins when carrying pQBR103 compared to plasmid to the plasmid free strain. All quantitative changes are significant to p <0.01 using a Mann-Whitney signed rank test.

4.4.2 Identification of proteins by Mass spectrometry analysis

4.4.2.1 Proteins further analysed

In Table 4.4 the 23 proteins that were further analysed are listed. These include all qualitative differences that were identified (19 in total over both conditions) and also the four proteins that were upregulated to the greatest extent when SBW25 with plasmid was grown on the R2A medium. Of the 23 proteins the vast majority of these where weak in intensity relative to the average protein spot intensity determined by 2-D gel electrophoresis. The 23 proteins are displayed in Figures 4.2 and 4.3.

4.4.2.2 Proteins identified by peptide mass fingerprinting

In total only three of the possible 23 proteins analysed showed homology to the pQBR103 genome. No proteins analysed showed homology to *P. fluorescens* SBW25 chromosome. The low level of identification is largely a result of a unknown sample contaminant that reduced the mass spectrometry signal (this is discussed in section 4.5).

Of the three protein samples only two could be unambiguously identified as being encoded on the pQBR103 plasmid genome, by using a combination of peptide mass fingerprinting and peptide sequencing (Table 4.5, proteins B2 and D3). These two samples represented the same protein expressed on both R2A and WA pea seed exudate confirming what was expected from the 2-D analysis. No peptide sequences were obtained for the third protein, D2. In itself the PMF homology is too low to confidently infer that the protein is encoded by pQBR103 CDS number 079 (chapter 2). However, further evidence is given by the unambiguous identification of the protein encoded by CDS 079 in a previous in house experiment (unpublished). Comparison of the isoelectric point and molecular size of the protein identified as CDS 079 previously and protein D2 in this study were congruent.

Table 4.5 summaries the mass spectrometry results for all three protein samples, B2, D3 and D2. By comparing the proteins identified in the context of the annotated plasmid sequence (chapter 1) the proteins could not be linked to any particular functional response, as they were located in regions of the genome/operons where no functional identification could be ascribed.

	Condition		Spot number*	Qualitative or Quantitative change	Spot strength	Trypsin concentration (ng μl^{-1})
a	**R2A**	**SBW25 + pQBR103**	A1	Qualitative	Weak	0.33
			A2	Quantitative	Weak	0.33
			A3	Qualitative	Intermediate	0.67
			A4	Qualitative	Weak	0.33
			A5	Qualitative	Intermediate	0.67
			A6	Qualitative	Weak	0.33
			A7	Qualitative	Weak	0.33
			A8	Qualitative	Weak	0.33
			A9	Quantitative	Weak	0.33
			A10	Qualitative	Weak	0.33
			A11	Qualitative	Weak	0.33
			A12	Quantitative	Intermediate	0.67
			B1	Quantitative	Weak	0.33
			B2	Qualitative	Strong	6.67
	R2A	**SBW25 Only**	C1	Qualitative	Weak	0.33
			C2	Qualitative	Weak	0.33
			C3	Qualitative	Weak	0.33
b	**WA PEA**	**SBW25 + pQBR103**	D1	Qualitative	Weak	0.33
			D2	Qualitative	Weak	0.33
			D3	Qualitative	Strong	6.67
	WA PEA	**SBW25 Only**	E1	Qualitative	Weak	0.33
			E2	Qualitative	Strong	6.67
			E3	Qualitative	Weak	0.33

Table 4.4. Table of the protein differences observed between SBW25 with and without plasmid that were grown on a) R2A media b) WA supplemented with 4% pea seed exudates. All proteins listed were further analysed by Mass spectrometry. All qualitative differences observed are represented in the Table. Only the four proteins with the largest expression increase in the presence of pQBR103 on R2A plates are shown. Concentration of trypsin used in digestion of each sample is given. * Spot numbers correspond to those shown in Figures 4.2 and 4.3.

Protein number	B2	D3	D2
Condition	R2A	WA + 4% pea exudate	WA + 4% pea exudate
Corresponding Figure	4.2b	4.3b	4.3b
PMF* result coverage, score	13%, 38	6%, 15	23%, 28
Peptide sequences	VSYNNVTLSGTATR, SINGNFQTR, AFANADEQQLAR	VSYNNVTLSGTATR, SINGNFQTR, AFANADEQQLAR	None
pQBR103 CDS number	134	134	079
Class of identified gene	Transmebrane	Transmembrane	Conserved hypothetical
Annotated function	Outer membrane autotransporter	Outer membrane autotransporter	Undetermined

Table 4.5. MALDI - TOF/TOF mass spectrometry results for three protein samples showing homology to the pQBR103 genome. * PMF = Protein Mass Fingerprint. Class of identified gene and annotated function is as concluded in chapter 2.

4.5 Discussion

4.5.1 Proteomic response is different on R2A media and pea exudate agar

In this chapter the expression of plasmid encoded proteins in response to different media, R2A and WA with pea seed exudate was investigated. In addition the effect of plasmid carriage on host protein expression was also studied.

Comparing total numbers of plasmid encoded proteins from both environments it was found that more were associated with R2A (n=10) compared to pea seed agar (n=3). Further analysis of these protein spots indicated that those expressed on pea seed agar where common to the R2A condition. This suggests these common protein spots either represent constitutively expressed proteins or proteins that are expressed in response to host i.e. not in response to the environment. Therefore, it is likely that pQBR103 offers little if any adaptive benefit to host on pea seed agar. In comparison, the specific induction of plasmid encoded proteins by the R2A medium suggests pQBR103 may offer some adaptive benefit to this environment. With regards to the effect of plasmid carriage on host protein expression, in both media conditions plasmid carriage was associated with a similar number of qualitative and quantitative changes. As some of these changes differed between the two media this suggests a degree of host plasmid interaction in response to environment.

4.5.2 Plasmid encoded protein expression

In this study it was hypothesised that more plasmid encoded proteins would be expressed in the pea seed condition compared to R2A. The rationale for this was that the former was predicted to most closely resemble the environmental conditions host and plasmid would experience in nature. The fact that the converse was found to be true is suggested to be a reflection of amount of exudate used and not necessarily because pQBR103 does not offer a benefit in seed germination.

In this study 4% total volume pea seed was used. The exact inorganic and organic constituents and their concentrations in the pea seed exudate were not determined.

Nevertheless other studies that have analysed particular constituents of pea seed varieties have shown it to be high in carbohydrates in particular simple sugars (Roberts et al., 1999; Roberts et al., 2000) and high in other sources essential to bacterial growth, such as amino acids (Lanfermeijer et al., 1992). At 4 % concentration this may represent a highly rich medium, unnatural compared to concentrations of exudates that Bacteria would experience in the natural environment. Consequently, it may be possible that at lower concentrations, when nutrients are more limited, that the plasmid may offer the host some fitness benefit.

Even though more potentially plasmid encoded protein spots where identified on R2A compared to pea seed agar, the fact that only ten were identified, of which seven were found to be specifically expressed on R2A, was still small in comparison to the predicted coding capacity of pQBR103. What was clear from the analysis of plasmid expression on both media was that plasmid expression is tightly regulated. This fits with biological theory that plasmid maintenance in a cell, and ultimately in the population relies on reducing metabolic burden in the absence of fitness benefits. Regulation of niche specific genes is essential to this reduction. So too is the regulation of plasmid copy number and plasmid functions, such as conjugative transfer, which is again a trade off between segregation stability and metabolic burden (Paulsson and Ehrenberg, 1998). Ultimately the success in reducing the metabolic cost depends on the environmental conditions and history of plasmid host co-evolution (Paulsson, 2002).

4.5.3 Plasmid and host expression, co-regulation

In this study one of the most notable findings was the degree to which pQBR103 carriage effected the expression of host encoded proteins. For both media conditions a similar number of significant qualitative and quantitative changes in host expression resulting from plasmid carriage was observed (pea seed n=40, R2A n=48). Comparing these changes they were found not all to be common to both environments i.e. some were environmentally specific responses. From the analysis of plasmid expression and plasmid directed host expression, two inferences can be made. Firstly, plasmid and host chromosome interact to give a global response to the environment. Secondly, for this to occur suggests a historical familiarity between plasmid and host.

In this study the detection of global responses and co-regulation between plasmid and host was not unusual or unexpected. Firstly, annotation of pQBR103 identified a number of regulatory genes which could putatively be suggested as having a role in host gene regulation (chapter 2). Secondly, global interactions between multiple independent replicons have been described previously in other bacteria (Ow et al., 2006; Guerrerio et al., 1998; Chen et al., 2000). However, what remains elusive in this study was the nature of the observed interactions between pQBR103 and host. In the aforementioned studies these interactions were shown to be symbiotic in nature. Without detection of the functional response of both plasmid and host it is not safe to assume this is also true for pQBR103 and *P. fluorescens* SBW25. It is possible these interactions were purely an antagonistic response.

4.5.4 Protein identification

In this chapter useful insights have been gained into the regulation and expression of pQBR103 proteins in different environments. Insights into the interaction of pQBR103 and *P. fluorescens* SBW25 has also been gained. However, the lack of protein identification has reduced the interpretation and conclusions that can be brought from the study. For example, only two putative pQBR0103 CDS can be confirmed by this investigation. Also, the nature of the plasmid response to R2A conditions cannot be determined; are these potentially plasmid encode spots reflective of a single operon or regulon response, or indicative a multiple responses that are not regulated co-ordinately?

However, it must not be assumed that if more proteins were positively identified that a greater insight into the biology of this plasmid would be gained, largely because of the large proportion of pQBR103 without an ascribed function. This point is illustrated by the two proteins identified, CDS079 and CDS134, whose identification has not led to a better understanding of the plasmid. Likewise, the regions of the plasmid genome identified by the IVET strategy (chapter 2) to be environmentally expressed has done little more to increase our understanding of the plasmid.

Several explanations can be given for the lack of positive identification of proteins in this chapter. Firstly, the weak concentration of the spots chosen for identification.

Secondly, the power to resolve proteins of similar pI and size. Thirdly, contaminants suppressing the mass spectrometry signal.

In this study SYPRO protein stain was used instead of coomassie blue as this is proven to be more sensitive (Berggren et al., 2000). Using SYPRO stain likely increased the number of proteins that could be detected by 2-D analysis. Most of the protein spots of interest in this study where at the cusp of the detection limit of this SYPRO stain. Even by pooling protein excisions from multiple gels, in house experience has shown the resulting pools are still at the limit of what can reasonably be expected to be identified by mass spectrometry. The problem of identification was further hampered by the practical consideration that in 2-D gel electrophoresis proteins of similar isoelectric points and molecular size overlap with each other making protein excision at best difficult, and at worst impossible. This is particularly true when using 11 cm 2-D gels (as used in this study) compared to the industry standard size of 24 cm.

For the reasons above, and common too many proteomic investigations, it was accepted that not all the proteins excised would be positively identified (in fact most wouldn't be). Nevertheless, it was reasonable to expect that some of the stronger proteins would be identified. In the process of protein excision and identification sample contamination is a common problem. This is usually an abundance of trypsin from protein digestion or keratin contamination (Ashcroft, 2003). In this study this was not found to be a problem but an unknown contaminant was. This contaminant effectively exacerbated the problem of identifying weak protein samples by suppressing the mass spectrometry signal and thus greatly decreased the sensitivity of detection.

4.6 Future directions

This study has represented a logical progression from the comparison of *P. fluorescens,* with and without plasmid on a rich laboratory medium, to more "natural" media, predicted to represent the conditions plasmid and host would experience in nature. In doing so this study has given a further insight into pQBR103 mediated response to differing environments. Nevertheless, the media used in this study was still artificial, which may explain the small level of plasmid expression identified. Therefore, based on these findings a future direction would be to move further towards more natural systems. As pQBR103 was initially isolated from the sugar beet phyllosphere (Lilley and Bailey, 1997a) and shown to impart fitness in this environment (Lilley and Bailey, 1997b), this logically would be to compare expression *in situ,* ideally on sugar beet leaves.

However, although in situ analysis would be the next progressive step from this study it has to be accepted that the conditions naturally faced by host and plasmid could never truly be recreated for proteomic analysis, for several reasons. Firstly, inoculated cells would have to be grown on sterilised leaves so that all proteins could be attributed to host or plasmid. Therefore, the complex community interactions on the leaf surface and their effect on host and/or plasmid would be lost. Secondly, it is unlikely that cells would grow at sufficient density on leaves to enable proteomic analysis. Instead cells would have to be grown in leaf plugs (inoculated bacterial pools on the leaf surface) or at worst grown on plant tissue agar. In both instances growth would be atypical compared to the characteristic micro-colonies formed by the bacterium in nature (Timms-Wilson, unpublished). Another consideration that would need to be addressed in future endeavors is protein identification. This would have to be substantially increased to make the project informative and financially justifiable. Dependent on cell yield detection may be increased by using larger 2-D gels for proteome separation. This would likely improve separation, the ease of spot extraction and protein yield (due to amount of protein that could be loaded on to the gel).

Regardless of increased protein identification, proteomics studies suffer from the proportion of the proteome that can be detected which is generally as little as 20 % (Ashcroft, 2003). This is largely because of the number of proteins that are below the detection limit of the stain, problems separating highly basic or acidic proteins (Volker and Hecker, 2005) and also because secreted protein will be absent from the analysis. Hence, the number of proteins that can potentially be identified is always going to be lower than the true number of qualitative and quantitative differences between the two experimental conditions. Consequently it would be ideal to extend this research to complement it with transcriptional analysis, such as using a *P. fluorescens* SBW25 pQBR103 genome microarray to investigate environmental response. By using transcriptional analysis the sensitivity of detecting the genomic response to the environment would be increased in comparison to proteomics. This has been demonstrated previously (Yoshida et al., 2001). However, as transcriptomics can only infer protein expression (as it cannot account for post transcriptional regulation and post translational modification) there are always going to be discrepancies between transcriptional patterns and *de facto* protein expression (Guerreiro et al., 1999). Therefore, it is proposed in future that both "omic" approaches are used in investigating plasmid and host response to the environment as their combined value is likely to be greater than the sum of their parts.

Chapter 5: The Plasmid Genome Database

www.genomics.ceh.ac.uk/plasmiddb/

5.1 Introduction

Bacterial genome sequencing has greatly increased scientific understanding of how bacteria adapt and evolve, and central to this is the role plasmids play in this process (as outlined in chapter 1). Nevertheless, without comparing a newly acquired genome, such as pQBR103, in the context of the wider genomic collection valuable insights which would otherwise be gained could be lost. Over the coming years plasmid genomes, like all genetic elements, will inevitably continue to be sequenced and deposited in the public databases (EMBL, GenBank, DDBJ). Although such databases provide an essential role in capturing sequencing data and making it available to the wider community, they provide little more than access to individual genome reports and make no attempt to assess and draw conclusions on the data as a whole.

In relation to all the plasmid genomes that have been completely sequenced the overall quality, limitations and therefore, ultimate usefulness of the collection is difficult to assess. In part this may be as a consequence of the genomic information associated with genome sequences (metadata) not being available to aid such assessments. However, a number of projects have sought to compare specific aspects of plasmid replicons or make metadata available for specific plasmid groups. Table 5.1 lists some of the more successful and relevant online plasmid projects and their corresponding locations.

Name of Database	Reference	Link/URL
Database of Plasmid Replicons (DPR)	Osborn. Unpublished	www.essex.ac.uk/bs/staff/osborn/DPR/DPR_ introduction.htm
Genome Database of Naturally occurring plasmids	Ong et al., Unpublished	www.biochem.ucl.ac.uk/bsm/PLASMID/main page.htm
Aclame Database	Leplae et al., 2004	http://aclame.ulb.ac.be/

Table 5.1. Online databases relating to the analysis of plasmid sequences

The first of these was the Database of Plasmid Replicons (DPR). The aims of the project were to facilitate rationalising the classification of bacterial plasmids as well provide an insight into plasmid evolution, based on comparison of replication proteins. It is note worthy as a premier database dedicated to plasmid replication and is a true community resource as it provides information for all *rep* types, not just of plasmids of specific interest to the curator. However, although it was hoped that this classification may be used as a framework with which to link other plasmid metadata, this goal unfortunately was never reached. Another site hosted at University College London is the Genome Database of Naturally Occurring Plasmids. At this site salient features of a number of genomes are collected including known incompatibility grouping and known phenotypes e.g. transfer and antibiotic resistances. However, the site is specialised (it hosts only 21 genomes of which most are from *E. coli* hosts) and so its usefulness is limited to only a subset of the plasmid scientific community. Furthermore, it does not summarise or enable interrogation of the metadata stored at the site.

Unlike the first two sites considered, the third, the Aclame database, is still actively being curated and is funded directly. The original aim of the Aclame database was to build a comprehensive classification of functional modules that make up all MGEs not just plasmids, on the premise that MGEs consist of shared modules and that the distinction into plasmid, transposon and phage are largely artificial (Toussaint and Merlin, 2002). More recently (after the PGD went live) the site displayed genomic metadata. This, however, is secondary to its original aims and does not go beyond what can be automatically obtained. Also, the metadata held at this site is not easily accessible and so is not available for comparison. Therefore, based on what is publicly available to the community, a database dedicated to the collection of genomic metadata is required. The design of a database that is populated by automatic and manually curated metadata will enable the most to be gained from the current plasmid collection that is publicly available. Futhermore it will provide a framework with which to link to other plasmid related data and informatics resources.

The first version of a plasmid genome "database" (PGD) was a spreadsheet created by manual searches of the NCBI website (www.ncbi.nlm.nih.gov/) and curated by Lars Molbak. In total 460 complete naturally occurring plasmid genomes were identified. These represented plasmids listed as completely sequenced genomes on the site and also plasmids that were mis-categorised at the NCBI website. To this early database a web interface was put on by the thesis author and published (Molbak and Tett et al., 2003).

Although, the manually parsed spreadsheet offered benefits to the immediate users it was limited in a number of ways. Firstly, manual parsing of GenBank document was time consuming and prone to human error. Secondly, there was no easy way to compare what information is captured in the spreadsheet compared to what is newly available at the NCBI site when updating the spreadsheet. Thirdly, plasmid data generated was restricted and not freely accessible and therefore, did not benefit from suggestions/error checking from the wider plasmid community. To minimise these shortcomings it would be beneficial to design a semi-automatic updating database to store plasmid genome metadata, ultimately this will enable pQBR103 to put in context of plasmids that have been completely sequenced.

5.2 Chapter Aims

The overall aim of this chapter is to put pQBR103 in context of all completely sequenced plasmid genomes and to give an assessment of the plasmid collection as whole. To do so it is necessary to create a web accessible and community responsive database dedicated to the capture and manipulation of plasmid genomic metadata.

This overall aim encompasses

Using data from existing public databases to:

1. Automatically collect, parse and store data of complete plasmid genomes and associated metadata (from genome submission reports) in the PGD
2. Make this data manageable and accessible to the wider plasmid community
3. Integrate community and curator modifications to this dataset and make these modifications visible to the user

Using data from existing public databases (Databases, Scientific publications, other online resources) to:

1. Go beyond genome submission reports to collect further metadata publicly available for each plasmid genome
2. Use publicly available annotation software to create additional features for each plasmid and make this available to the wider community

This will enable the following hypothesis/questions to be addressed.

In Chapter 2 it was hypothesised that the large percentage of orphan genes (60%) on pQBR103 represented genes that are of ecological relevance that are hitherto undescribed. In this chapter the question posed is whether the high percentage of orphans predicted reflects how unusual/unique pQBR103 is to other large phytosphere plasmids and those from other environments?

5.3 Materials and Methods

5.3.1 Design of the PGD

All scripts for the database were written in Perl/Perl-Cgi in a text editor on a LINUX operating system, unless otherwise stated. The PGD website was created in HTML with the aid of Dreamweaver 4 software (Macromedia, UK), or manually in a text editor. Meta data generated was stored in a MySQL database.

The database created can be viewed at www.genomics.ceh.ac.uk/plasmiddb/
A schematic of the design of the database and flow data within it is given in Figure 5.1. Essentially the flow and creation of user accessible metadata can be divided into five parts:

1 New plasmid genomes released are identified at various locations at the NCBI site using a Monitor utility program. Newly released genomes and updated GenBank versions of existing plasmid genomes are then downloaded from the NCBI site using the script, "Get updates".

2 For each of the plasmid genomes that are held locally (these include the mis-catergorised plasmids at the NCBI site and those downloaded from the plasmid genome page of NCBI) are then automatically parsed for nine metadata fields (outlined in Table 5.2).

3 The extracted features from the plasmid GenBank files are modified automatically. These modifications are a combination of parsing errors due to atypical GenBank genome reports, updates or errors in the metadata feature values that were introduced at the point of sequence submission to the public databases (NCBI, EMBL or DBJ). These modifications are made by the curator or by suggestions from the community. The automatic merging of the modified metadata with the GenBank extracted metadata prevents modifications made to PGD being overwritten in updates to the database.

4 Both the modifications list and the modified metadata features are uploaded into a MySQL database.

5 The plasmid genome metadata can be accessed by the user via a web accessible interface on the PGD. All metadata features are available for sorting and searching. Additionally every plasmid genome entry is linked back to NCBI.

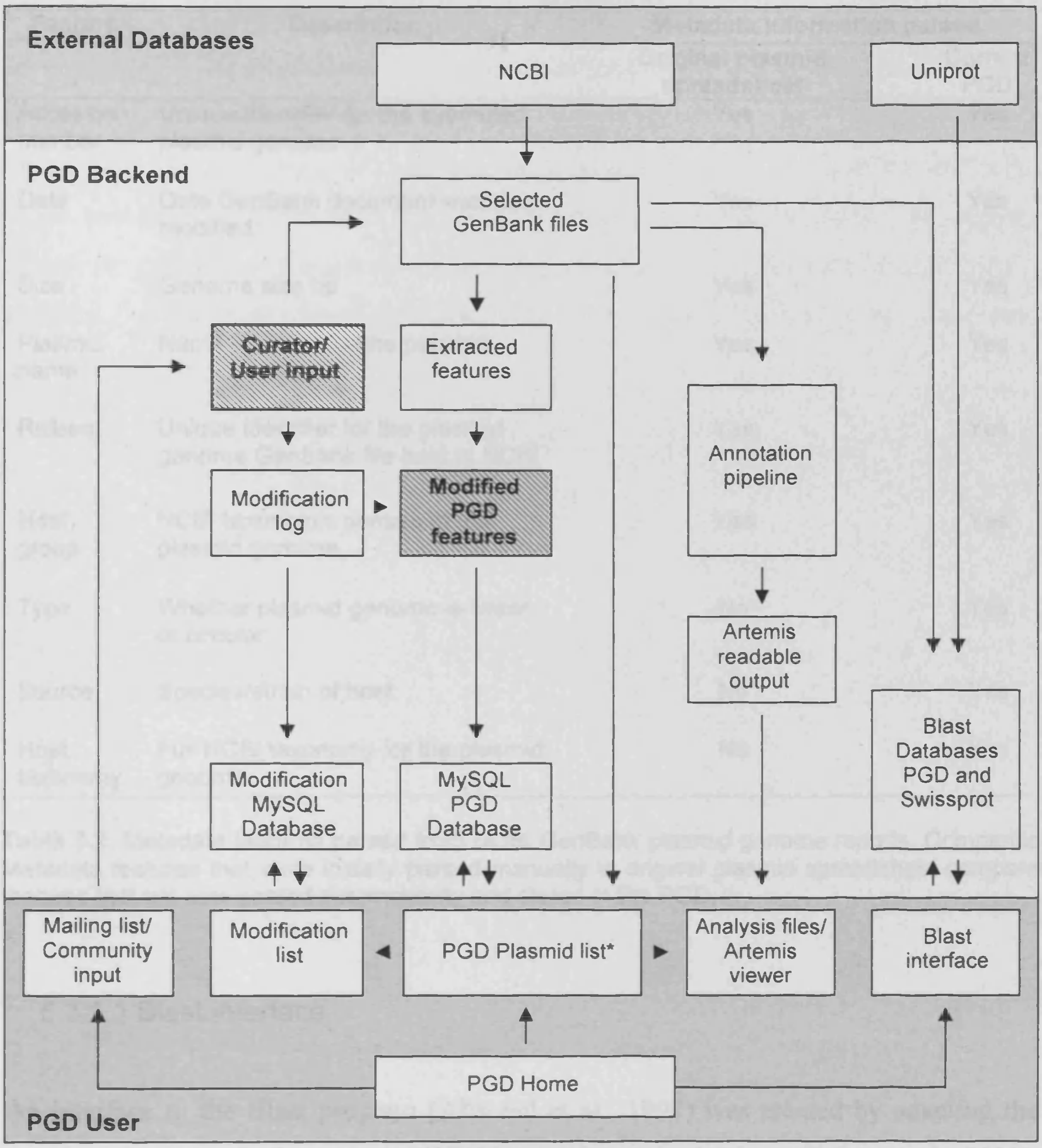


Figure 5.1. Simplified schematic of the flow and interaction of data in the PGD. Hatched boxes refer to curator and community input that is central to the PGD project. * "The Plasmid list" links to external resources held at NCBI i.e. taxonomy and genome reports.

Feature	Description	Metadata information parsed	
		Original plasmid spreadsheet	Current PGD
Accession number	Unique identifier for the submitted plasmid genome	Yes	Yes
Date	Date GenBank document was last modified	Yes	Yes
Size	Genome size bp	Yes	Yes
Plasmid name	Name if known for the plasmid	Yes	Yes
Refseq	Unique identifier for the plasmid genome GenBank file held at NCBI	Yes	Yes
Host group	NCBI taxonomic domain of the plasmid genome	Yes	Yes
Type	Whether plasmid genome is linear or circular	No	Yes
Source	Species/strain of host	No	Yes
Host taxonomy	Full NCBI taxonomy for the plasmid genome	No	Yes

Table 5.2. Metadata features parsed from NCBI GenBank plasmid genome reports. Comparison of Metadata features that were initially parsed manually in original plasmid spreadsheet compared to features that are now parsed automatically and stored in the PGD.

5.3.1.1 Blast interface

An interface to the Blast program (Altschul et al., 1997) was created by adapting the interface from the *P. fluorescens* SBW25 encyclopaedia database (Spiers et al., 2001). A custom nucleotide database of all plasmids constituting the PGD was made available for searching. Additionally a current version of the protein Swissprot database is available for searching on the PGD site.

5.3.1.2 The first pass annotation pipeline

Annotation software freely available to the wider scientific community was employed in the analysis and annotation of the pQBR103 genome (chapter 1). This software was wrapped to form a first pass annotation pipeline, the results of which were parsed so they can be viewed in the Artemis annotation viewer (Rutherford et al., 2000). A brief list of software included in this pipeline is given in Table 5.3. The parsing scripts used in this pipeline were written by a number of developers, including the author, at CEH Oxford. These parsing scripts became a precursor to the YAMAP project (www.bioinformatics.org). This pipeline was used to analyse all the plasmids held in the PGD, enabling pQBR103 to be compared to other plasmid in the PGD. The analysis results for each plasmid genome can be viewed on the PGD, if the user has JavaSE 5.0 runtime environment (www.java.sun.com), by using Artemis java web start (www.sanger.ac.uk/Software/Artemis).

PGD Software and Database searches	Brief description	Reference / URL
tRNAscan-SE	Software designed to detect ribosomal transfer RNA genes in DNA sequences	Lowe and Eddy, 1997 http://selab.janelia.org/tRNAscan-SE/
MSATfinder	Software used to identify microsatellite repeats in DNA sequences. (Microsatellites are short direct repeats of 1-6bp in size)	CEH Oxford www.genomics.ceh.ac.uk/msatfinder/
Emboss einverted repeats	Identifies inverted repeats in nucleotide sequences (sequence of nucleotides that has a reverse complement further down stream) that could be indicative of stem loop structures	Rice et al., 2000 http://bioweb.pasteur.fr/docs/EMBOSS/einverted.html
Emboss palindrome repeats	Identifies inverted repeats in nucleotide sequence, that fall within a certain repeat size range and distance from each other.	Rice et al., 2000 http://bioweb.pasteur.fr/docs/EMBOSS/palindrome.html
Embosss etandem	Identifies tandem repeats in a nucleotide sequence	Rice et al., 2000 http://bioweb.pasteur.fr/docs/EMBOSS/etandem.html
Glimmer2	Glimmer is a software package designed to predict coding sequences (CDS regions) in microbial genomes.	Delcher et al., 1999 http://cbcb.umd.edu/software/glimmer/
BigBlast	Software that randomly dissects each sequence into 1 kb pieces and blasts them against a user specified database. In the PGD this is against GenBank db (see below)	Sanger Centre www.sanger.ac.uk/Software/ACT/BigBlast/

Table continued overleaf…..

PGD Software and Database searches	Brief description	Reference / URL
TranstermHP	Software for identifying rho-independent transcriptional terminators	Kingsford et al., (Publication in preparation) http://cbcb.umd.edu/software/transterm/
Pfam database	The CDS predictions made by Glimmer are searched against Pfam database. Pfam database is a databse of two parts, Pfam-A represents over 8000 curated protein families, Pfam-B contains a large number of small families.	Pfam reference: Finn et al., 2006 www.sanger.ac.uk/Software/Pfam/
NCBI nr Genbank database	A genetic sequence database of all publicly available DNA sequences	Benson et al., 2006 www.ncbi.nlm.nih.gov/Genbank/

Table 5.3. Software and Database searches used in the third party analysis pipeline. For each sequence analysis program or database search a description is given. The respective reference for each piece of software/database is given, where available, as well as a link to that software/database.

5.3.2 Manually curated fields; Environment of plasmid isolation

Additional fields that went beyond what was available for parsing from the GenBank documents i.e. characteristics of plasmid isolation, were curated manually. This information was collected by searching primary sequence publications, other relevant publications, sequencing centre websites responsible for that sequence submission and the NCBI genome project metadata collection (www.ncbi.nlm.nih.gov/). Each plasmid was categorised as animal associated or non-animal associated based on the environment from which the plasmid was either endogenously or exogenously isolated, if this could be determined. Animal associated environments include isolation from animal pathogens, commensals, endosymbionts, etc as well as from animal waste products e.g. a plasmid isolated from animal faeces was categorised as animal associated.

For non-animal associated plasmids above 50 kb in size more information regarding the environment/host they were isolated from was captured. Firstly plasmids were separated based on whether they were isolated from an aquatic (saline or fresh water) or from a terrestrial environment, if this could be resolved. For plasmids isolated from terrestrial environments, these were further separated based on whether they were isolated from the phytosphere, and if so, whether they were from a plant pathogen or from a host with a different plant association e.g. symbiotic. If not phytosphere associated, whether the plasmid had a known, or putatively annotated, polyaromatic xenobiotic degradation phenotype, or was isolated from a heavily xenobiotic contaminated environment was recorded.

5.3.3 Comparison of the large phytosphere plasmids to bacterial secondary chromosomes

Secondary chromosomes were downloaded from NCBI for all the completely sequenced Eubacterial genomes with multiple chromosomal replicons (as of 25th September 2006). For a genome of multiple chromosomal replicons, the secondary chromosome (or chromosomes where chromosome number was greater than two) were defined on the criteria of size only i.e. the primary chromosome was the largest. Of the 29 secondary chromosomes and 30 Eubacterial phytosphere plasmids greater than 50 kb, which have been included in the PGD, the presence of tRNA in each genome was identified using the default settings of tRNAscan-SE (Lowe and Eddy, 1997), as part of the annotation pipeline. Comparisons between the plasmids and secondary chromosomes were made on size and presence of tRNAs.

5.4 Results

As the aim of this chapter was to compare pQBR103s physical and metadata attributes to the plasmids of the PGD as a whole, pQBR103 was not included in the overall analysis of the plasmid collection to avoid comparisons to self.

5.4.1 Overall patterns observed in the PGD and putting pQBR103 in context of the plasmid collection

Since the PGD became live (October 2003) constant updates have indicated a linear trend in the number of plasmid genomes that have been deposited in the public databases, and thus captured by the PGD. The growth of plasmid genomes in comparison to the growth of microbial genomes completely sequenced and deposited in the public databases is shown in Figure 5.2. Currently the rate of plasmid genomes becoming available is greater than complete microbial genomes (11 per month compared to 6.5 per month). As of September 2006 the total number of plasmid genomes captured in the PGD stood at 929. The vast majority of this number (92%, n=852) comprised plasmids isolated from Eubacteria. The other 77 were isolated from Eukaryotic and Archaeal hosts. As pQBR103 was isolated into a Eubacterial host, these are considered further.

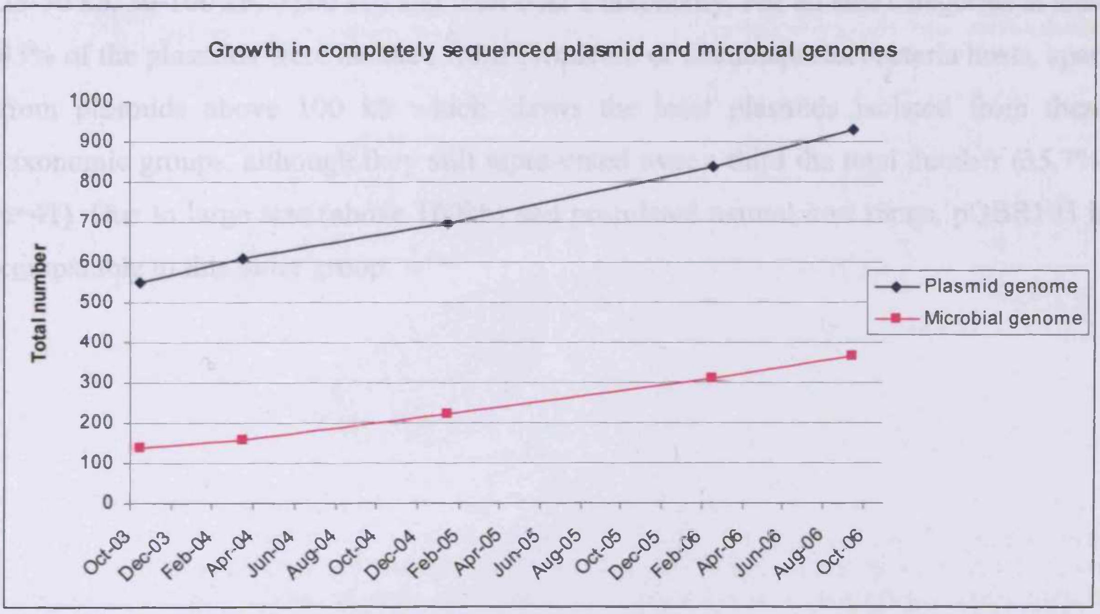
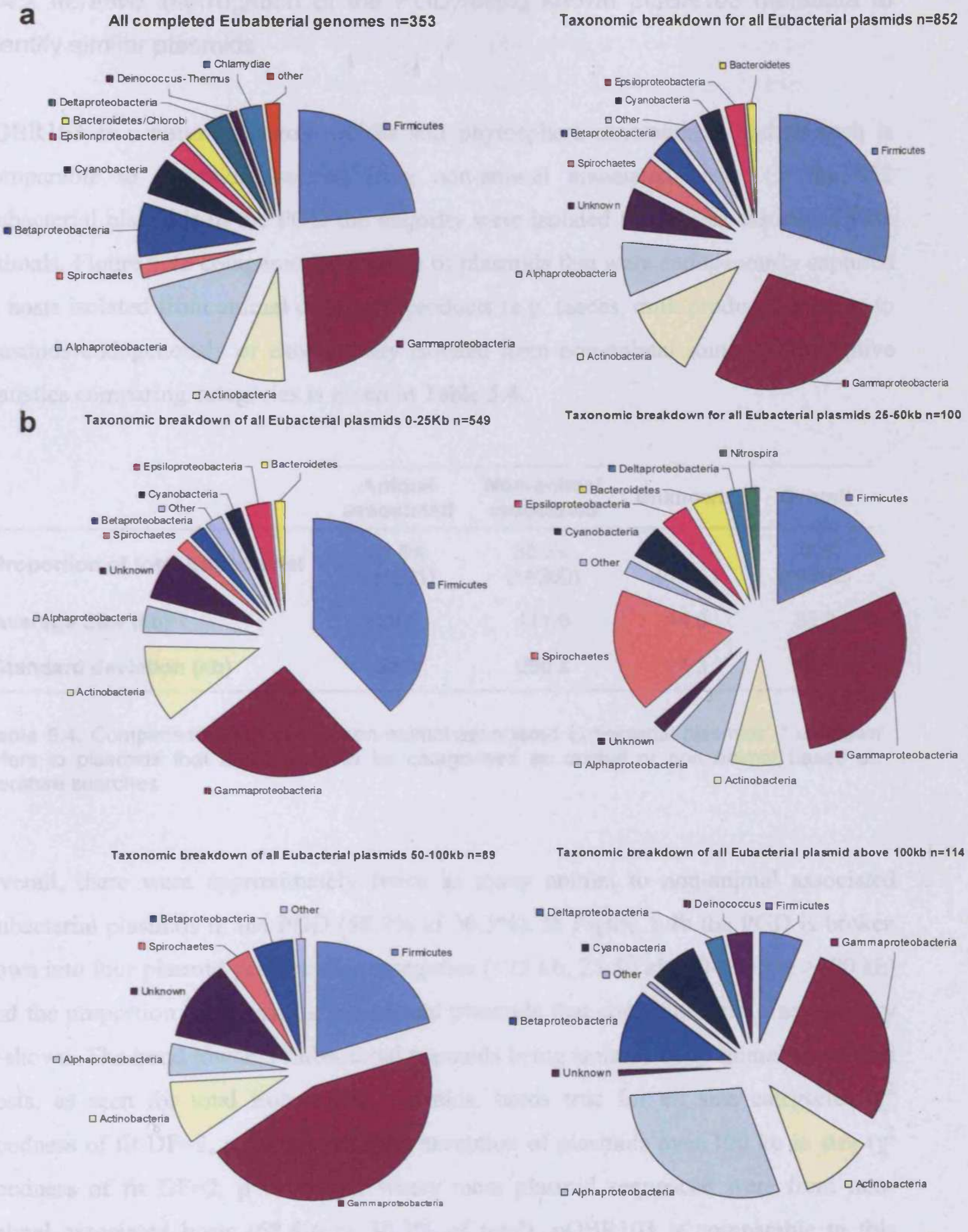


Figure 5.2. Growth in completely sequenced plasmid and microbial genomes. The pink line illustrates the linear growth of the total number of completely sequenced microbial genomes publicly available, since October 2003, at NCBI. The blue line illustrates the linear growth of the total number of completely sequenced plasmid genomes in the PGD since October 2003. Based on these Figures the average rate in growth is 11 and 6.5 genomes per month for plasmid genomes and microbial genomes respectively.

5.4.1.1 pQBR103 in relation to plasmids from hosts of different bacterial phyla

pQBR103 is a large plasmid known to exist and replicate in *Pseudomonas sp* of the Gammmaproteobacteria, and was sequenced independently of host. Of the Eubacterial plasmids in the PGD, in total 231 plasmids of the 852 Eubacterial plasmids in the database came from 117 whole Eubacterial genome projects. This represented 27.1% of the total number of plasmids in the PGD. The remaining 72.9% (n=621) were sequenced independently of the host chromosome. In Figure 5.3a the relative proportion of whole Eubacterial genomes from NCBI and the Eubacterial plasmids in the PGD isolated from different taxonomic phyla and Proteobacteria is compared (Due to the number of Proteobacteria these are separated to the level of class). In the whole organism Eubacterial genomes the largest number were attributed to Firmicutes and Proteobacteria (in particular the Gammaproteobacteria class). As of September 25th 2006 (n=353), 49.0% (n=173) were from Firmicutes or Gammaproteobacteria. Similarly a large proportion of plasmid genomes populating the PGD were also isolated from Firmicute and Proteobacterial hosts (collectively 58.0%, n=494).

Figure 5.3b shows a breakdown of the Eubacterial plasmids into size categories (<25 kb, 25-50 kb, 50-100 kb, >100 kb) and their host's taxonomy. For all size categories at least 45% of the plasmids were isolated from Firmicute or Gammaproteobacteria hosts, apart from plasmids above 100 kb which shows the least plasmids isolated from these taxonomic groups, although they still represented over a third the total number (35.7%, n=41). Due to large size (above 100kb) and postulated natural host range, pQBR103 is comparable to this latter group.



All completed Eubabterial genomes n=353

Taxonomic breakdown for all Eubacterial plasmids n=852

Taxonomic breakdown of all Eubacterial plasmids 0-25Kb n=549

Taxonomic breakdown for all Eubacterial plasmids 25-50kb n=100

Taxonomic breakdown of all Eubacterial plasmids 50-100kb n=89

Taxonomic breakdown of all Eubacterial plasmid above 100kb n=114

Figure 5.3 Plasmid genomes in the PGD separated based on host's NCBI taxonomy. **a**, All whole organism Eubacteria genomes (NCBI 29th September 2006) and all Eubacterial plasmids in the PGD separated at the taxonomic level of phyla of host, except for proteobacteria that are separated at the level of class. **b**, Taxonomic breakdown of Eubacterial plasmid hosts in the PGD at the level of class for different plasmid genome size subsets. Other indicates bacterial classes that are either singularly represented, or constitute less than 1% of the total plasmids for each individual pie chart.

5.4.2 Iterative interrogation of the PGD, using known pQBR103 metadata to identify similar plasmids

pQBR103 is a natural plasmid of the soil phytosphere environment and as such is comparable to plasmids isolated from non-animal associated hosts. Of the 852 Eubacterial plasmids in the PGD the majority were isolated from hosts associated with animals. Figure 5.4a compares the number of plasmids that were endogenously captured in hosts isolated from animal or animal products (e.g. faeces, milk products) relative to plasmids endogenously or exogenously isolated from non-animal sources. Descriptive statistics comparing categories is given in Table 5.4.

	Animal associated	**Non-animal associated**	**Unknown***	**Overall**
Proportion of total Eubacterial	58.8% (n=501)	30.5% (n=260)	10.7% (n=91)	100% (n=852)
Average Size (kb)	30.1	111.9	14.0	53.3
Standard deviation (kb)	51.4	230.8	24.3	139.1

Table 5.4. Comparison of animal to non-animal associated Eubacterial plasmids. * unknown refers to plasmids that are unable to be categorised as animal or non animal based on literature searches

Overall, there were approximately twice as many animal to non-animal associated Eubacterial plasmids in the PGD (58.8% cf 30.5%). In Figure 5.4b the PGD is broken down into four plasmid genome size categories (<25 kb, 25-50 kb, 50-100 kb, >100 kb) and the proportion of animal to non-animal plasmids that constitute each size category is shown. The trend towards Eubacterial plasmids being isolated from animal associated hosts, as seen for total Eubacterial plasmids, holds true for all size categories (χ^2 goodness of fit DF=2, p <0.05) with the exception of plasmids over 100 kb in size (χ^2 goodness of fit DF=2, p <0.0001), where most plasmid sequenced were from non-animal associated hosts (68.4% cf 30.7% of total). pQBR103 is comparable to this group. In Figure 5.5 the distribution of genome size of plasmids isolated from animal compared to non-animal associated hosts is given. As shown there are a number of outliers above the 90th percentile, this was particularly true for plasmids from non-animal associated hosts). pQBR103 at 425 kb represents the 17th largest plasmid genome to be completely sequenced in context of the PGD, and would fall into this group of outliers. In comparing the two populations, non-animal associated hosts were significantly larger than those from animal associated hosts (Mann-Whitney test, p <0.0001), with an average genome size three times as large (Table 5.4).

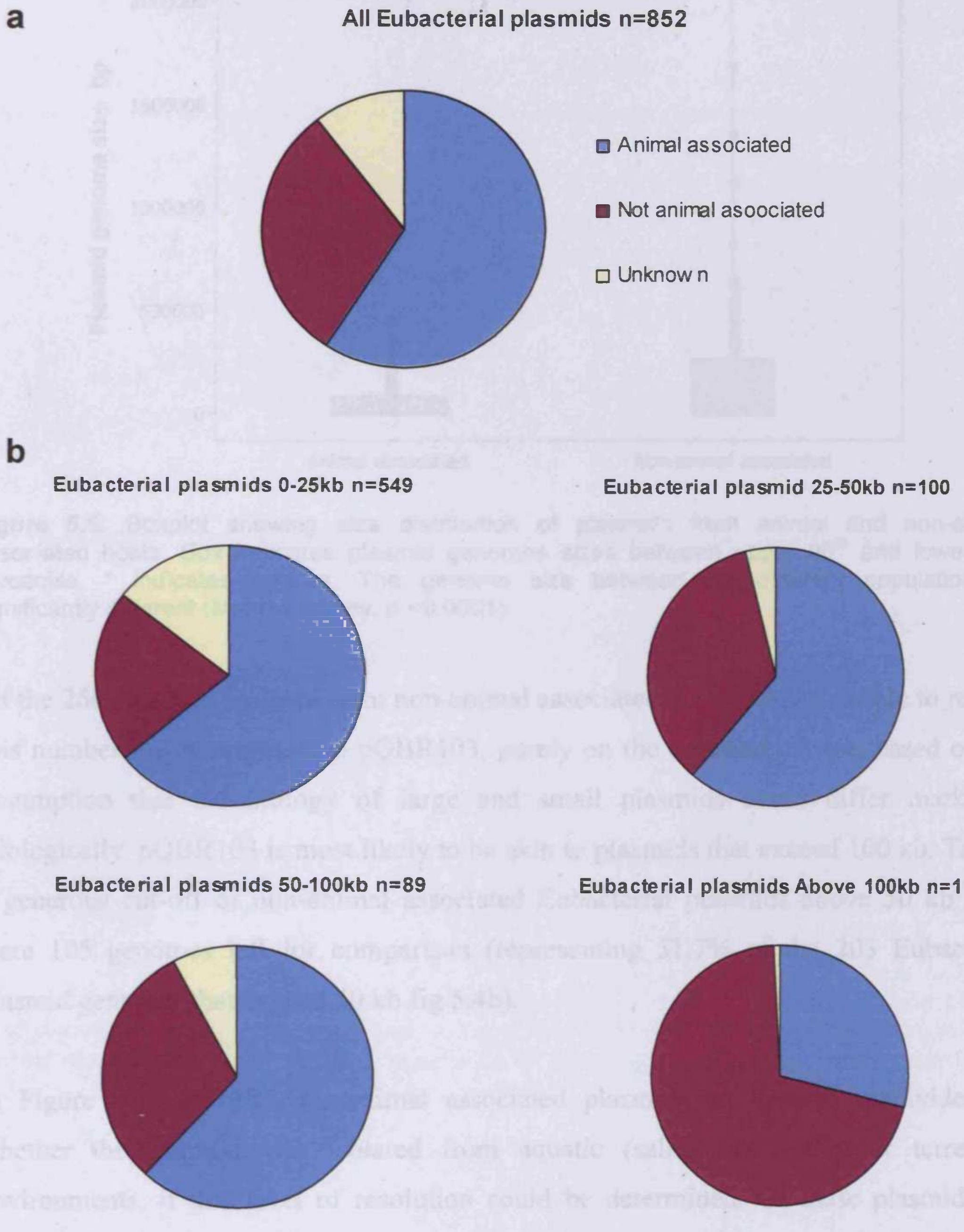


Figure 5.4 Distribution of plasmids isolated from different animal or non animal associated environments. Colours: **Blue**, animal associated; **Maroon**, Non animal associated; **Yellow**, unknown. Fig 5.3a The total number of Eubacterial Plasmids in the PGD isolated from the different environments if known. Figure 5.3b Isolation distribution of Eubacterial plasmid subsets from the PGD based on size ranges. The distribution of plasmids above 100kb significantly differed from that of the total Eubacterial plasmids in the PGD (X^2 goodness of fit DF=2, $p < 0.0001$). Other plasmid size categories did not significantly differ (X^2 goodness of fit DF=2, $p < 0.05$).

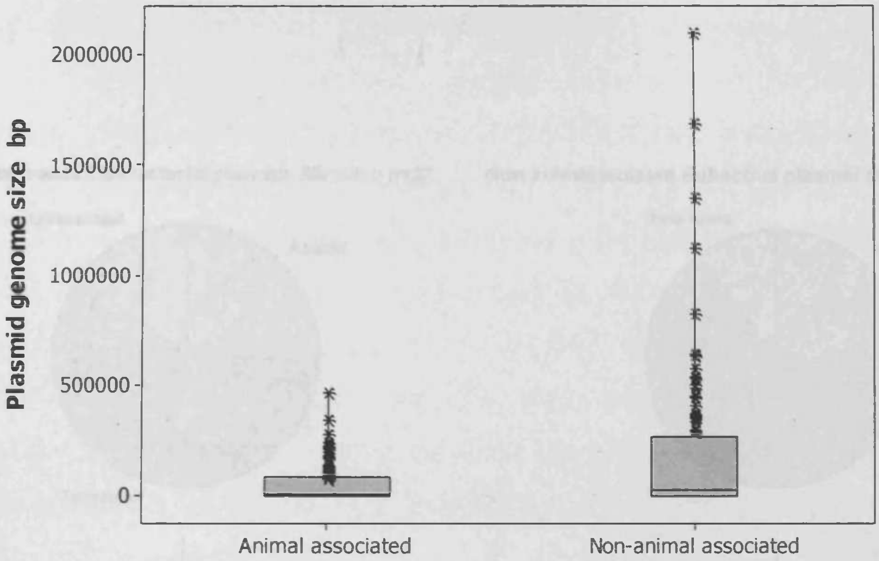


Figure 5.5. Boxplot showing size distribution of plasmids from animal and non-animal associated hosts. Box indicates plasmid genomes sizes between upper 90th and lower 10th percentile. * indicates outliers. The genome size between the different populations is significantly different (Mann-Whitney, $p < 0.0001$)

Of the 260 plasmids isolated from non-animal associated hosts it was possible to reduce this number for comparison to pQBR103, purely on the criterion of size, based on the assumption that the biology of large and small plasmids could differ markedly. Biologically, pQBR103 is most likely to be akin to plasmids that exceed 100 kb. Taking a generous cut-off of non-animal associated Eubacterial plasmids above 50 kb there were 105 genomes left for comparison (representing 51.7% of the 203 Eubacterial plasmid genomes that exceed 50 kb fig 5.4b).

In Figure 5.6a the 105 non-animal associated plasmids are further subdivided on whether the plasmid was isolated from aquatic (saline and fresh) or terrestrial environments, if this level of resolution could be determined. Of these plasmids the proportion from terrestrial environments, akin to pQBR103's isolation environment was found to be over twice as large as those from aquatic environments (n=67 vs n=26, unresolved; n=13).

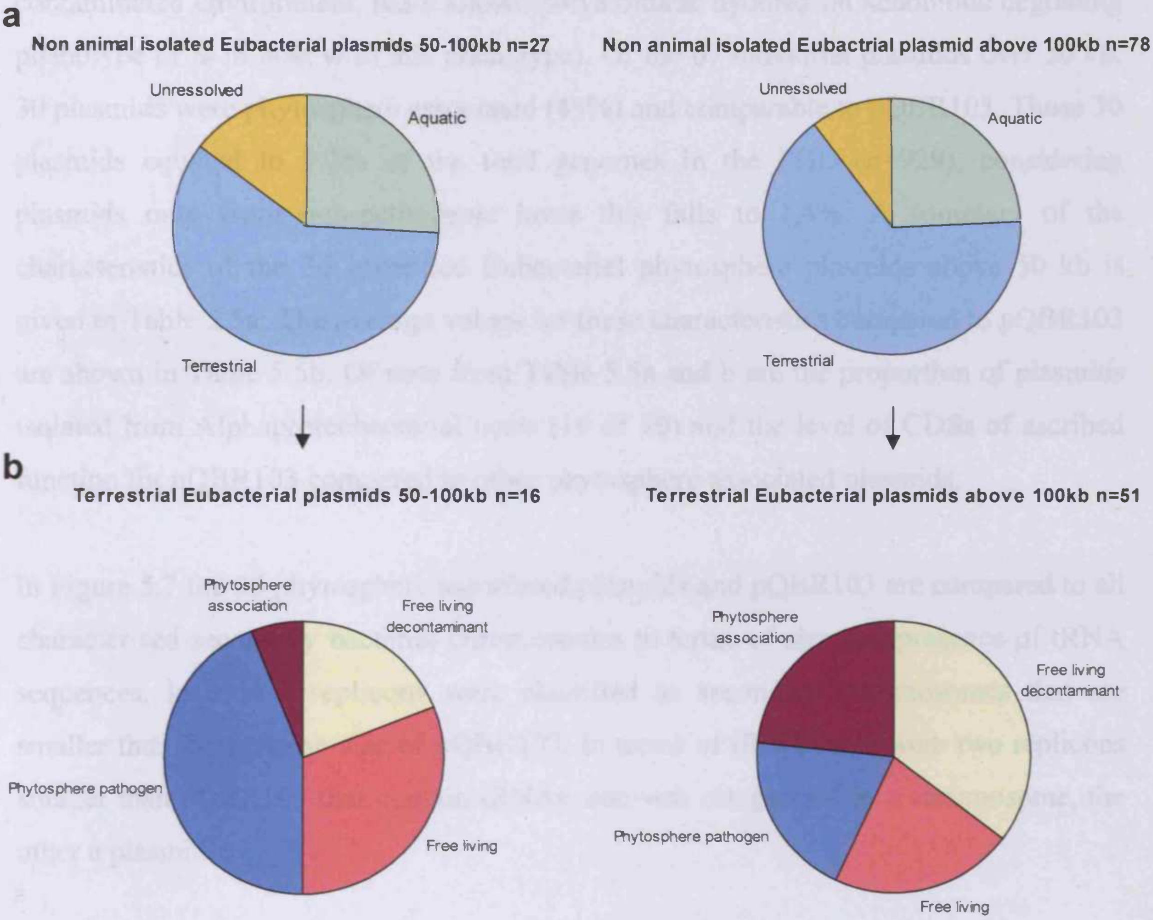


Figure 5.6 Iterative breakdown of non-animal associated Eubacterial plasmids into location either exogenously or endogenously isolated a. separation to aquatic and terrestrial environments b. further resolution of the environment of isolation for terrestrially isolated Eubacterial plasmids.

In Figure 5.6b terrestrial data for the 67 plasmids is further subdivided into four categories. These categories are, phytosphere associated, phytosphere pathogen, free living and free living decontaminant (plasmid isolated in a heavily xenobiotic contaminated environment, has a known polyaromatic hydorcaron xenobiotic degrading phenotype or is in host with this phenotype). Of the 67 terrestrial plasmids over 50 kb, 30 plasmids were phytosphere associated (45%) and comparable to pQBR103. Those 30 plasmids equated to 3.2% of the total genomes in the PGD (n=929); considering plasmids only from non-pathogenic hosts this falls to 1.4%. A summary of the characteristics of the 30 identified Eubacterial phytosphere plasmids above 50 kb is given in Table 5.5a. The average values for these characteristics compared to pQBR103 are shown in Table 5.5b. Of note from Table 5.5a and b are the proportion of plasmids isolated from Alphaproteobacterial hosts (19 of 30) and the level of CDSs of ascribed function for pQBR103 compared to other phytosphere associated plasmids.

In Figure 5.7 the 30 phytosphere associated plasmids and pQBR103 are compared to all characterised secondary bacterial chromosomes in terms of size and presence of tRNA sequences. In total 3 replicons were classified as secondary chromosomes that are smaller than the genome size of pQBR103. In terms of tRNA there were two replicons smaller than pQBR103 that contain tRNAs, one was categorised as a chromosome, the other a plasmid.

	Phytosphere pathogen	Phytosphere associated	Total Phytosphere
Number of Plasmid genomes	17	13	30
Number unique strains	12	6	18
Average Genome Size kb (s.d)	293.1 (488.3)	500.1 (486.6)	382.8 (490.3)
Average % protein coding ± (s.d)	81% (0.06%)	82% (0.07%)	81% (0.06%)
Average coding density, gene kb^{-1} ± (s.d)	0.97 (0.14)	0.94 (0.11)	0.96 (0.12)
Average % CDS unascribed function* (s.d)	35.2% (14.8 %)	35.8% (13.3%)	36.2% (13.6%)
Average CDS size bp (s.d)	844 (110)	891 (95)	860 (104)
Number of plasmids with tRNA	2	3	5
Eubacterial classes (number)	γ-protoebacteria (9) α–proteobacteria (7) β–proteobacteria (1)	α–proteobacteria (12) Exogenous (1)	γ-protoebacteria (9) α–proteobacteria (19) β–proteobacteria (1) Exogenous (1)

Table 5.5a.

	Total Phytosphere average	pQBR103
Size kb (s.d)	382.8 (490.3)	425.1
Percentage protein coding (s.d) ±	81% (0.06%)	83%
Coding density, gene kb^{-1} (s.d) ±	0.96 (0.12)	1.12
Percentage CDS unascribed function* (s.d)	36.2% (13.6%)	80%
Average CDS size bp (s.d)	860 (104)	738
tRNAs	Yes (5/30)	No

Table 5.5b

Table 5.5. Comparison of genomic characteristics of phytosphere plasmids a) Genomic characteristics of phytosphere associated plasmids greater than 50 kb in the PGD. b) Comparison of phytosphere plasmids to pQBR103. s.d refers to standard deviation ± Coding density and percentage coding are values taking from published annotation and do not include pseudogenes. * Percentage CDS unascribed refers to published annotation. Percentage given includes both hypothetical and orphan CDS.

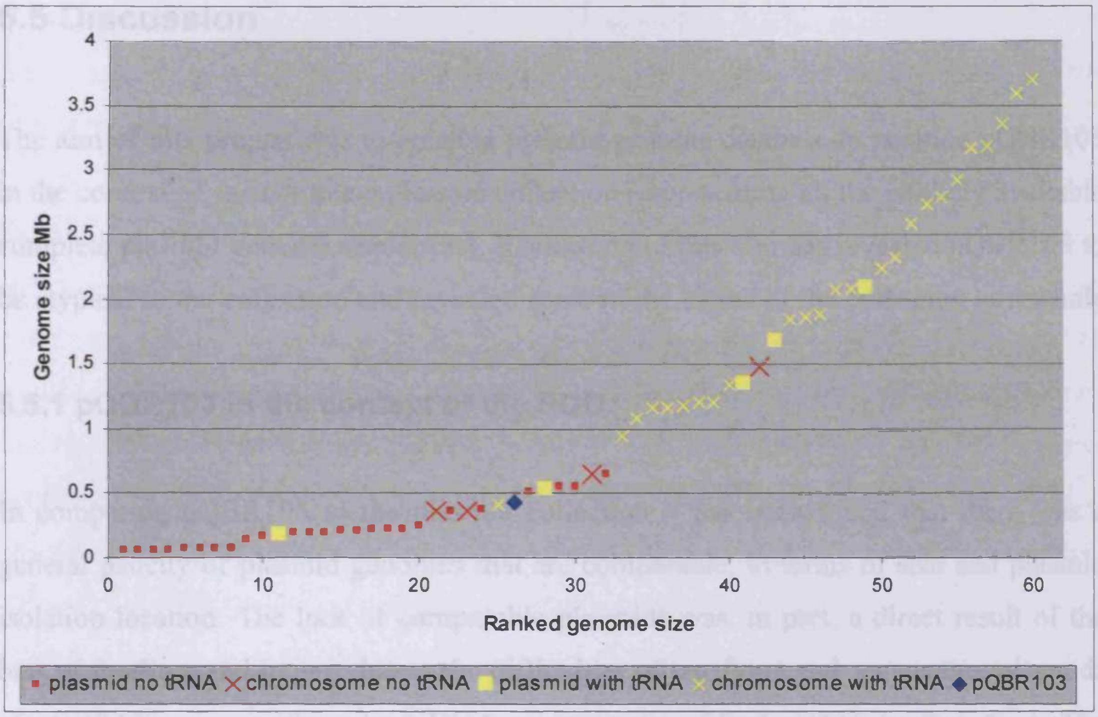


Figure 5.6. Comparison of Phytosphere associated plasmids greater than 50 kb and all Eubacterial secondary chromosomes. Square data points refer to plasmid genomes, crosses refer to secondary chromosomes. Yellow indicates tRNA identified on that replicon, and red no tRNAs identified using tRNAscan-SE with default settings.

5.5 Discussion

The aim of this project was to create a plasmid genome database to position pQBR103 in the context of the complete plasmid collection (representing all the publicly available complete plasmid genome sequences). Realisation of this aim has revealed pQBR103 to be atypical to the collection and revealed some of the biases of the collection as a whole.

5.5.1 pQBR103 in the context of the PGD

In comparing pQBR103 to the plasmid collection it has been found that there was a general paucity of plasmid genomes that are comparable, in terms of size and plasmid isolation location. The lack of comparable plasmids was, in part, a direct result of the bias of funding, and in part due to the difficulties of purifying and sequencing plasmids of pQBR103s size, without host DNA contamination (discussed in next section). The paucity of information with which to compare pQBR103 may offer an explanation for the large proportion of orphan CDS and CDS with unascribed function (Chapter 2). The plasmids that were most akin to pQBR103, in terms of size and being non-pathogenic phytosphere associated, were those of the Rhizobiales (in total 12). Six of these plasmids were attributable to the single *R. etli* CFN 42 genome project (Gonzalez et al., 2006). In comparison to these plasmids, pQBR103 had a noticeably higher proportion of CDS of unascribed function (80% compared to ~33% averagely for the rhizobia). This discrepancy was likely due to the biotechnological and agronomic importance of the rhziobia. This has attracted much investment into understanding the plant microbe interactions of this group (as reflected by the multiple genomes determined for the rhizobia). This investment both *in silico* as well as in empirical functional approaches has greatly increased the number of genes which can be putatively ascribed a function. This is in contrast to pQBR103 which represents the only sequenced member of a group of mercury resistant phytosphere plasmids among which functional investigation is still in its infancy.

A further explanation of the high level of non-functionally assigned CDS could be because pQBR103 host range is unknown. Although there is evidence to suggest pQBR103s natural host range includes Pseudomonads, whether it persists and associates with non-culturable hosts in the natural habitat cannot be determined. Therefore, it is at least possible that pQBR103s novelty is because it has gained sequences from a genetic pool that is largely unstudied due to limitations in culturing.

Probably one of the most notable features of pQBR103 in comparison to the plasmid collection as a whole is its size. Plasmids isolated from non-animal associated hosts were significantly larger than those in animal association, and pQBR103 was not an exception. This observed size difference may be explained by the heterogeneity and level of selection pressure. This is weak and diverse on free living and, to a lesser extent, transiently animal associated hosts, compared to the narrow and steep selection faced in animal associated hosts, and it is this that tolerates a larger genome size (Doolittle, 2002; Turner et al., 2002). However, this increase in genome size, associated with a free living lifestyle, is postulated to be specialisation to a niche or a small subset of niches, and not a generalist mechanism offering fitness to a disparate range of environments. (as discussed in chapter 1). Over time, based on the local adaptation hypothesis of Eberhard (1990), plasmids collect suites of fitness related genes that cumulatively increase fitness to that environment. As a consequence, genome size increases. For some plasmids that form close associations with a particular host this can potentially lead to the accumulation of "housekeeping genes". It is at this point the plasmid/secondary chromosome boundary becomes blurred. This blurring has been observed for a number of replicons associated with the phytosphere. For example, the *A. tumefaciens*'s secondary linear chromosome (Goodner et al., 2001; Wood et al., 2001) and *Ralstonia eutropha* H16 secondary chromosome (Accession: AM260480) which were classified as secondary chromosomes yet both have plasmid replicon systems suggesting a plasmid origin. Further evidence of this blurring, and secondary chromosomes having plasmid origins is provided by comparisons of reciprocal homologous gene content between large plasmids, secondary chromosomes and primary chromosomes. This has shown secondary chromosomes to be more akin to plasmids than primary chromosomes based on gene content (S Turner unpublished, personal communication). Therefore, the question arises, is pQBR103 a secondary chromosome?

Based purely on replicon size pQBR103 was at the "cusp" of the plasmid / chromosome "divide" (Figure 5.7). However, size alone cannot be offered as evidence of being a secondary chromosome. Further support that pQBR103 is not a secondary chromosome is that it doesn't contain any recognised, essential "housekeeping genes". Also, that it is not limited to a particular host spp and has observed seasonality in the field population (Bailey et al., 2001)

5.5.2 The plasmid collection is biased

Nearly one thousand plasmids populated the PGD, as of September 2006. In comparing pQBR103 to this collection (the primary aim of this chapter), it has revealed the collection to be far from representative of the natural plasmid diversity found in the environment, and was skewed in several different respects. Concentrating on plasmids from the Eubacterial domain (although the findings are likely to be applicable to other domains), the first of the biases was the disproportionate sequencing of plasmids from particular taxonomic groups.

By far the largest proportion of plasmids in the PGD was associated with Proteobacteria and Firmicute hosts. This was paralleled by the number of bacterial genomes that have been completed for these groups, despite most plasmids being sequenced independently of host. At some level it could be argued that both bacterial and plasmid genomes coming for these taxonomic divisions is a reflection of their relative natural abundance in the environment. For example, although different proportions of these phyla have been determined (using culture independent methods) in different animal and non-animal associated environments (Janssen, 2006; Wang et al., 2005; Dekio et al., 2005), in general these phyla represent a large proportion of the total bacterial community relative to other phyla. Therefore, it would follow that a higher proportion of genome sequences should be expected compared to those from less abundant phyla. However, even if this is accepted, there are always going to be taxonomic biases because of varying degrees of culturability between and within different phyla.

Historically genome sequencing was reliant on culturable bacteria, because of the amount of DNA material that is required for the sequencing process. With advances in technology such as amplifying sufficient DNA from single bacterial cells for sequencing (Zhang et al., 2006), this situation may, in part, become alleviated (Hutchison and Venter, 2006). Moreover, metagenomic projects are also enabling, in some cases, the elucidation of almost complete bacterial genome sequences (Venter et al., 2004). Nevertheless, the current genome collection was derived from the culturable proportion of bacteria. Estimates have suggested that as little as 1% of environmental bacteria are culturable using current techniques (Sharma et al., 2005). With regards to animal associated bacteria this proportion is higher, in part, due to investments in developing new culture techniques. Yet, regardless of the exact proportion of bacteria that are culturable, the abundance of phyla that constitutes this proportion is not reflective of the total abundance found in nature, with some phyla being highly recalcitrant. For example, the degree of culturability within firmicutes and proteobacteria is quite high relative to Acidobacteria, where only a small fraction is culturable (Quaiser et al., 2003). Yet, in some soil environments this phylum can constitute a large dominant proportion of the bacterial community (Fry 2000; Quaiser et al., 2003; Janssen, 2006) As sequencing to date has largely been restricted to the small proportion of the total bacterial diversity that is culturable, it is fair to conclude that the Eubacterial genome collection was inevitably going to be biased. However, is it fair to conclude that plasmids from particular taxonomic groups are over represented in the PGD? Two considerations must be made in answering this. Firstly, plasmids vary in their host range and secondly, they can be exogenously captured. Therefore, potentially a number of plasmid in the collection could at least offer adaptive advantage and be associated with non-culturable bacterial hosts. Obviously there is no way of telling how many plasmids sequenced were endogenously isolated that can, and do, naturally exist in non-culturable hosts. However, the number that were exogenously isolated can be determined. Based on this low Figure alone, it is likely that the taxonomic bias of sequencing is as true for plasmids as it is for chromosomal replicons.

The bias towards the sequencing, and indeed research into, the culturable proportion of the Eubacterial domain was only in part a result of choice; mostly it was accepting the practical limitations of culturing. Where choice introduces the largest bias into both the plasmid, and the bacterial, genome collection it was in sequencing genomes associated with a particular lifestyle (e.g. pathogens) or from a particular environment (contaminated land). This selection process obviously further skews the number of genomes associated with particular taxa. By far the vast majority of plasmids that populate the PGD were animal associated, representing animal pathogens, commensals etc (Figure 5.4). Again, this bias is synonymous to the situation that was recognised in the bacterial genome collection.

The bias towards sequencing plasmids from animal associated hosts may be explained in a number of ways. Firstly, there was a bias in the scientific community towards funding sequencing of plasmids from animal associated hosts, particularly from human pathogens, as well as plasmids of industrial/biotechnological importance. This was because of implications to public health and economic drivers. Secondly, as plasmid genome sizes in non-animal associated hosts are larger, there were practical limitations, including cost and isolating large plasmid DNA away from host at sufficient quality/purity for sequencing (a non trivial task). Although the second explanation was likely to impact on the bias of the database, the first explanation is probably most accountable. The question is whether this is an historical bias or not?

With regards to plasmid genomes, information regarding the number currently being sequenced and the justification for sequencing is unknown. In contrast this information is available for bacterial genomes. As has been shown in this chapter, the biases between bacterial genome and plasmid genome sequencing are largely synonymous, therefore it is likely valid to extrapolate observations from bacterial genome projects and apply them to the plasmid genome projects. Probably the most reliable project summarising the overall features of microbial, and in particularly bacterial genome projects, is the Genome OnLine Database (Gold) (Kyrpides, 1999; Liolios et al., 2006). As of June 2006 there were 1248 ongoing bacterial genome projects, of which 78% were funded on the basis of being either of biomedical or biotechnological importance. In comparison, only 17% were selected as of environmental relevance. These Figures are interesting in two ways. Firstly, as of September 15th 2006 there were only 353 bacterial genomes sequenced and deposited in the public databases, compared to the 1278 (~four times the number) in progress as of June that year. This illustrates the enormous growth in genomes that are soon to become publicly available. Secondly, the bias towards animal associated genome projects still exists, and is not simply an historical artefact. However, accepting the bias towards animal related sequencing to be “the norm”, at least in terms of total numbers there is going to be increase in plasmid genomes sequenced that are of environmental relevance. For example, there is a project dedicated to the sequencing of approximately 100 environmental soil plasmids that are to be deposited in the public databases in the coming years (Eva Top, personal communication). Such projects on the large scale are sure to offer large scale comparative analyses that will in turn offer a greater understanding of the behaviour and impact of these plasmids in natural environments.

5.5.3 Future challenges

The creation of the PGD provided an invaluable tool to understanding pQBR103 in the context of the complete plasmid collection. From an early stage the PGD has been made available to the wider scientific community to enable a resource from which to do similar analyses. As of September 2006 the PGD had received in excess of ten thousand credible hits to the database interface i.e. a better indicator than the PGD home page as it shows entering of the site. A benefit of making this resource publicly available is that by correspondence to the author the community has greatly enriched and error checked the site. At this point all the aims of this chapter have been met. However, some consideration must be given to the future challenges faced by the PGD, and the plasmid community as a whole, regarding the collation and storage of plasmid sequences and associated metadata.

5.5.3.1 Community and collaboration

The most important aspect of a database such as the PGD is its curation and management, as it must be kept up to date to be useful to the wider community. However, generally online databases associated with plasmid analysis are created by the "hobbyist" in their spare time. Consequently, and no matter how well meaning at the point of conception, they can become mothballed with dead links in cyberspace. Therefore, it is important that the data is collected and shared, not in competition, but in collaboration. The PGD has been instrumental in that process. For example, historically the PGD was the only site to collect metadata and make this accessible to the community. It is refreshing to see that the PGD has prompted other sites to adopt similar strategies e.g. the Aclame (Leplae et al., 2004). Also, NCBI now provides plasmid genome summary data e.g. the number of CDS, percentage coding and interactive genomic maps; features that were not available at the conception of the PGD. In addition, the curated PGD is an integral part of the GenomeMine database project (www.genomics.ceh.ac.uk/cgi-bin/genomemine/gminemenu.cgi), which brings together metadata for all small genomes e.g. whole Eubacterial genomes, viruses, phage etc.

Although, the community is recognising the importance of metadata collection resources are still disparate. Faced with the growth of plasmid genomic sequences, a more formal community effort is required if the most is going to be gained from data and this is considered next.

5.5.3.2 Growth of data

In the current age with the development of technology one of the most challenging problems in science, across all disciplines is the growth, manipulation and storage of data. This is particularly true to genomic sequencing. Over the last three years the growth in both Eubacterial genome and plasmid genome sequences in the public databases has been linear. However, over the same time the growth in Eubacterial genome projects that are in progress has grown at exponential rate (source: GOLD). There are also plans to sequence all validly named species in the ATCC culture collection (representing approximately 6000 genomes) over the next 10 years (G. Garrity, personal communication Editor of Bergey's Manual of Systematic Bacteriology). Thus, the number of Eubacterial genomes being published is set to increase. This is likely to be synonymous with plasmid genomes. In this chapter to pull any meaning from the plasmid collection required manual input, i.e. determination of the isolation origin of the plasmid and inferred lifestyle of that plasmid (e.g. pathogensis, symbiosis, xenobiotic decontaminant). Nevertheless, despite considerable manual searching at every stage of the iterative process, more plasmids were labelled unknown, and fell out of the analyses. For example, some plasmids could not even be categorised as animal or non-animal associated plasmids. This reliance on manual data collection is inadequate and problematic. Firstly, because this is time consuming and laborious to the curator and may be unmanageable in the face of the number of genomes becoming available. Secondly, there is always going to be a limitation on the data that can be extracted when going beyond the original submission report, so this will make many plasmids useless with regards to making any assessment of the overall plasmid collection.

The problem of metadata collection is not restricted to plasmid genomics but faced by all genomic scientists interested in small genomes (Field and Hughes, 2005). In the case of Eubacterial genomes this paucity of information has largely been filled by gathering the information manually from Bergey's Manual (Garrity., 2001). However, the situation is worse for plasmids and genomes other than bacterial chromosomes as there is no Bergey's Manual equivalent. Nevertheless, despite a Bergeys Manual, lots of strain specific and other associated data is lost at the point of submission that greatly reduces the richness of the data set. Increasing and standardising the information obtained at submission has been considered (Field and Hughes, 2005; Ward et al., 2001). Currently there is a community led response to agree on the information that can reasonably be expected to be captured at submission for all Eukaryotic, Archaeal and Eubacterial genomes and metagenomes (Field et al., in review). This standard is akin to other data standards like MIAME (Brazma et al., 2001) for microarray experiments. By the community complying to the Minimal Information about a Genome Standard (MIGS), it is hoped more meaningful comparisons can be brought between different genomes, including plasmids. The PGD project is supportive and committed to this community initiative. Additionally, pQBR103s genome, sequenced as part of this thesis, was used as a test case for standardising the data regarding plasmid genomes. Table 5.6 shows the data that is to be captured based on MIGS checklist for plasmids as of September 06.

5.5.3.3 Automatic analyses and mining plasmid collections

Community response to data capture at the point of submission is vital if the most is to be made of the ever increasing plasmid genome dataset. However, this is only one aspect of enriching our data, and comes with limitations. The other aspect is mining the sequence data itself. Capturing information from the community that does not change is important, such as the plasmid isolation mechanism and location, as well as experimentally verified phenotypes. Where problems occur, it is in comparison between inferred genome features, i.e. comparisons between annotations. This is problematic in two ways. Firstly, the process of annotation is not standardised at the outset and so is influenced by the annotators. This was illustrated by the discrepancies given in the two independent sequencing and annotations of the same pTiC58 plasmid sequence. Here, the number of published functionally unascribed genes differs almost three fold between the two annotations (Goodner et al., 2001; Wood et al., 2001). Secondly, and inevitably, the original annotation becomes outdated as annotation tools become more accurate and more sequence information is deposited in the public databases. This is not to say that annotation is pointless, far from it, but that the annotation can be augmented by automatic data-mining such as the first pass annotation pipeline of the PGD, which provides this updated information directly to the user. Automatic annotation or analysis comes with its own limitations, but it does have the advantage that it allows uniformed comparison across data collections, as well as updating annotations in light of new functional assignments and additional gene sequences deposited in the public databases.

The power of combining manual and automatic analyses has been recently demonstrated by the Aclame project (Leplae et al., 2006). Here global comparisons were made between 500 completely sequenced plasmid genomes, in terms of gene and functional content. Due to high recombination rates and no 16S/18S rRNA gene equivalents, such comparisons afford a way of organising and relating plasmid genomes. However, with better metadata collection the true value of such analyses will become clear. In this thesis going beyond the metadata that was collected at the point of submission afforded an assessment of the current plasmid collection a whole. The logical step, therefore, would be to combine such metadata features, such as plasmid isolation location, with data coming from the Aclame project to see if gene content is correlated with plasmid isolation location and host lifestyle i.e. what are physical and functional attributes of plasmids isolated from different environments and from hosts of different lifestyle.

INVESTIGATION	
Study	
ORGANISM	
Complete genetic lineage (below lowest rank of NCBI taxonomy)	A group I plasmid according to Lilley and Bailey., 1996 criteria
Number of chromosomes (Virus: number of segments)	1 circular plasmid (sequence)
Estimated size (prior to sequencing)	330kb
Reference for the description of the biological material sequenced (isolate, soil sample etc)	Lilley and Bailey., 1997a
Information on access to the isolate sequenced	Available from CEH, Andrew Lilley / Mark Bailey
Source material identifiers: (cultures of micro-organisms: identifiers for two culture collections, specimens (e.g. organelles and eukarya): voucher condition and location)	None
Specific Host	Unidentified
Host specificity/range	Expected to include but not be restricted to *Pseudomonas* spp
Specific source of sample	Capture in a sugar beet field into a marked *Pseudomonas fluorescnes* SBW25
Encoded traits like antibiotic resistance	Mercury resistance, UV resistance
Incompatibility Group	Unknown
ENVIRONMENT	
Geographic location (latitude, longitude, depth / altitude of sample)	University Farm, Wytham, Oxfordshire
Date and time of sample collection	Unspecified
Habitat type	Sugar beet (*Beta vulgaris*) field. Agricultural
Environment (Bac/Arch:biotic = host, or abiotic; plasmids:medical, environmental, plant etc; same for hosts)	Environmental (Phyllosphere)
SAMPLE PROCESSING	
DNA preparation (DNA extraction method and amplification *e.g.* MDA, emPCR, plones)	Wheatcroft and Williams isolation (1981). Further enriched and purified by agarose electrophoresis and gel extracted using phenol and ethanol precipitation
Library Construction (library size, number of clones sequenced)	4,508 paired end-reads from one pUC19 library with insert sizes of 2.0-4.0 kb, and 357 paired end-reads from a second library with inserts of 1.4-2.0 kb
Sequencing Technology Used (e.g. Dideoxysequencing, Pyrosequencing, Polony)	Dideoxysequencing
Assay	
DATA PROCESSING	
Assembly (assembly method, estimated error rate and method of calculation)	phred/ phrap
Finishing strategy (status *e.g.* complete or draft, coverage, contigs)	Complete, 8.64-fold coverage

Table 5.6. Information captured for pQBR103 to make it MIGS compliant. (Field et al., in review). MIGS checklist is accurate as of September 2006.

Chapter 6: Discussion

6.1 Thesis findings

The pQBR plasmid collection represents of one of the best studied plasmid communities from a single geographical location to date. Of this collection, pQBR103 is by far the best characterised plasmid. The fitness of this plasmid has been demonstrated *in planta* (Lilley and Bailey, 1997b). To assess the determinants of this fitness and better understand the impact of this plasmid on host populations in the phytosphere, *in vivo* expression technology has been employed (IVET) (Zhang et al., 2004a;b). Although useful, this approach has given only limited benefits to the understanding of this plasmid and the pQBR collection as a whole. To complement and gain further understanding of the pQBR plasmids, in this thesis the full nucleotide complement of the pQBR103 plasmid was determined. Obtaining this sequence has fallen short from the analogy of cracking a piñata (whereby all its contents become evident), as no putative functional role could be ascribed to 80% of the pQBR103 CDS. For this reason alone, pQBR103 remains largely enigmatic. However, the sequence and subsequent genomic analyses it has afforded has certainly driven the understanding of these pQBR plasmids forward.

The occurrence and diversity of plasmids in naturally existing bacterial communities, both in aquatic and terrestrial environments, has been well documented. Selected examples include the river epilithon (Fry and Day, 1990), marine sediments and bulk water/air interphases (Sobecky et al., 1997; Dahlburg et al., 1997), xenobiotic contaminated soils and sludges (Khesin and Karasyova, 1984; Top et al., 1994), and of relevance to this thesis, the plant phytosphere. Here examples include the rhizobial pSym plasmids (Young and Wexler., et al 1988), *P. syringae* pathovar plasmids (Bender and Cooksey, 1986; Sundin et al., 1994) and the pathogen associated plasmids of Xylella (Hendson et al., 2001). What is clear from these studies is that plasmids are ubiquitous in natural environments. However, while plasmid genome sequencing and genomic characterisation of rhizobial plasmids and plasmids from contaminated environments are relatively mature, due to economic drivers, plasmid genomic sequencing and characterisation from other natural environments is still in its infancy. Therefore, while the occurrence of large phytosphere plasmids, such as the pQBR collection, may in itself not be unique in nature, genomic sequencing from this environment is, or at the very least unusual as shown by the analysis of the PGD collection (chapter 5). For example, inorganic mercury resistance has either been used as selection criteria or found to be common to many plasmids in natural environments (Khesin and Karasyova, 1984; Dahlburg et al., 1997). However, only 2% of the plasmids in the PGD could be ascribed putative inorganic mercury resistance.

A further testament of the uniqueness of sequencing from natural environments, and indicting that this represents a largely untapped genetic pool, was reflected in the amount of the pQBR103 genome that could not be ascribed a function (80% of CDSs). Based on previous fitness studies, *in vivo* analysis, *in silico* analyses and the local adaptation hypothesis of Eberhard (1990), it was predicted that pQBR103 encodes a higher proportion of genes that are related to adaptation in the phytosphere in comparison to the host chromosome. The annotation of the pQBR103 genome has provided an insight into the putative replication and maintenance functions of this plasmid, whereby a putative active partitioning system has been identified, a putative dimer resolution system and a credible minimal replicon. By contrast, very few ecologically relevant determinants could be identified; those identified included, inorganic and organic mercurial resistance, UV resistance and genes that may have a potential role in chemotaxis. All these functions have been implicated to contribute to host fitness in the phytosphere (Sundin and Murillo, 1999; Pandya et al., 1999; Bailey et al., 2001; de Weert et al., 2002). Disappointingly, however, they add little to what was already generally known for this plasmid genome. Having said this, a notable finding from analysis of pQBR103 was the identification of a number of putative regulators. These may be involved in host chromosome/plasmid global responses to perturbtions in local environmental conditions. Support that pQBR103 may be involved in these responses was provided by the findings of the proteomic studies (chapter 4). On two different environmental conditions plasmid carriage directly altered the qualitative and quantitative (>2 fold up or down regulation) expression of host encoded proteins. This host expression change was different between the two environments suggesting this may be a possible coordinated response by plasmid and host to the environment, as has been demonstrated for other hosts/plasmids (Guerreiro et al., 1998; Chen et al, 2000; Ow et al., 2005). This also suggests a historical familiarity between pQBR103 and pseudomonas hosts. Nevertheless, despite the findings from annotation and proteomic studies, the fitness imparted by pQBR103 remains largely cryptic.

One advantage afforded by genomic sequencing is the application of comparative genomics. In this thesis a PCR survey and a pQBR103 genomic array were used to estimate the diversity of the pQBR group I plasmids, and give insight into the population genetics of the group. With regards genetic diversity it was found that other group I plasmids were largely synonymous or subsets of pQBR103. In total, 5 plasmids of similar size to pQBR103 were indistinguishable from pQBR103, based on the PCR surveys (chapter 2). Two other smaller plasmids, pQBR44 and pQBR47, could be distinguished by the PCR surveys and the plasmid genomic array (chapter 2 and 3), but are predicted to share their entire nucleotide complement with pQBR103. Although, using only a single genome to investigate genetic diversity within the group I plasmids had its limitations, in so far as it could not identify novel sequences, based on the findings it suggested sequencing pQBR103 had accessed a large proportion of the genetic complement of this group. Therefore, the elusiveness of ecological function applied to pQBR103 can be extrapolated to the group I plasmids as a whole. Of particular interest from these studies were the inferences regarding group I genetic structure and intragroup mixing from the pQBR44 and pQBR47 results.

Plasmids largely fit the modular paradigm; whereby the genome can be defined as consisting of a "core" or "backbone", encoding functions related to plasmid maintenance, and regions containing traits beneficial to host adaptation. Whereas, the core is highly conserved amongst related plasmids, considerable variability is observed in the adaptive regions, usually mediated by IS and transposable elements (Thomas, 2000). In chapter 2, annotation of pQBR103 revealed plasmid maintenance functions were not tightly clustered to a particular region, and there was little evidence of IS and transposable elements, suggesting pQBR103 does not fit this modular model and might be largely recalcitrant to recombination. However, this latter point was contradicted by the genomic array analysis where considerable intragroup variability was observed, although the exact mechanism of it remains elusive. Although a highly clustered plasmid "backbone", by the definition of Thomas (2000) was not observed, the PCR surveys and genomic array experiments did reveal a region of commonality within the group I pQBR plasmids. This region was hypothesised to constitute the entire pQBR44 plasmid genome. What was interesting regarding this region were the putative functions that were not ascribed to it. For example, the *oriV* defined for the group I plasmid, pQBR11 (Viegas et al., 1997), UV resistance and most notably both candidate regions that may be involved in conjugative transfer (chapter 2). This latter point is of most interest, as other than broad spectrum mercury resistance, replication and self conjugation, these are the only defined phenotypes of pQBR44 to date.

The identification of divergent type IV conjugative systems in other interrelated groups of environmental plasmids, indicating independent acquisitions by horizontal gene transfer, has been described, for example in the *P. syringae* pPT23A plasmids (Zhao et al., 2005). However, as it was predicted that the common pQBR group I region constituted the entire pQBR44 genome, it follows that this region contained the genes relating to conjugative transfer. The fact that no identifiable conjugation system could be ascribed to this region suggests it might encode a highly divergent or novel system, and further exemplifies how unusual these plasmids are in context of current understanding of plasmid biology.

In addition to testing intragroup variability in this thesis, pQBR103 was also used to estimate intergroup variabiliy. At the Oxford field site a number of different groups have been identified, three of which have been shown to persist over successive growing seasons (group I III and IV). By comparing pQBR103 to representative members of these groups it was found that, other than the *mer* operon, pQBR103 is genetically distinct from these groups, suggesting these plasmids have independent evolutionary histories. Based on previous ecological studies it has been shown that these groups are not host specialists and that their occurrence is correlated with stage of sugar beet maturation. Therefore, it was hypothesised that these plasmids are partitioned by their specialisation to niche i.e. they are responding to different plant stimuli, and as such these plasmids represent complementing rather than competing groups. This is quite intriguing; the local adaptation hypothesis predicts that plasmids will accumulate functions related to adaptation to the local environment and this appears to be the case for the pQBR plasmids. For some plasmids that become closely associated with host, this can lead to the acquisition of functions essential to housekeeping. As such they set off down the path of becoming secondary chromosomes. An example where this has likely occurred are some of the rhizobial "plasmids" e.g. the pSymB of *S. meliloti* (Finan et al., 2001). Although the pQBR plasmids are large, and based on inferences from this thesis, highly specialised, in the author's opinion this does not provide evidence that these plasmids are on the path to becoming secondary chromosomes.

Firstly, based on endogenous isolations these plasmids were not found to be host type specific and secondly, they were only seasonally associated with hosts. In order to become essential to host survival i.e. become a secondary chromosome, plasmids must accumulate genes of "housekeeping" functions that are simultaneously lost by the host. The transient nature that has been observed for the pQBR plasmids suggests that the chance of this occurring is low.

What is clear from the annotation of pQBR103 and the inferences that can be made to the pQBR group I as a whole, is just how unusual these plasmids are in the context of other plasmids/plasmid groups that have been sequenced and described. Much of our knowledge is based on the study of particular *inc/rep* types, most of which are represented by the Couturier probes (Couturier et al., 1988). However, plasmid populations from natural environments have often failed to hybridise to these probes (Sobecky et al., 1997; Dahlberg et al., 1997), including the plasmids isolated from the sugar beet at the Oxford field site (Kobayashi and Bailey, 1994). Such findings, and the annotation of pQBR103 reveal just how under sampled and how poorly understood plasmids in natural environments are.

6.2 Future work Developments

Obtaining the pQBR103 sequence has afforded the first glimpse, at the genomic level, into the large mercury resistant pQBR phytosphere plasmids, moving the study of these plasmids into the genomic era. While much has been gained by annotating pQBR103 and subsequent genomic analyses, this largely marks the starting point from which to better understand the phytosphere importance of these plasmids.

In addition to future work given in the experimental chapters, given time, the immediate progression from this thesis would be to use gene disruption/knockout studies to confirm putatively ascribed functions for pQBR103/pQBR group I plasmids. Such targets include the putative stability and replication associated sequences, but perhaps of more interest is the region(s) responsible for conjugative transfer. To identify transfer related sequences, working with pQBR44 would be favoured due to its small size compared to other pQBR group I plasmids. While confirming the putative functions ascribed by annotation is a worthwhile exercise, this will not identify the determinants responsible for environmental fitness of these plasmids in the phytosphere. Therefore, it may seem the next logical step would be to conduct mutational studies and screen for loss of fitness *in planta* using micro/mesocosm studies. However, such studies are likely to be limited, as environmental fitness is unlikely to be attributable to a single plasmid genomic loci, but rather the combined effect of a number of disparate traits responding to a heterogeneous environment (Rainey, 1999), as was indicated by IVET experiments. Although useful, IVET studies had their limitations, a major one being they could only look at possible transcription at early plant growth stages (seedlings) (Zhang et al., 2004a :b). For pQBR103 it was demonstrated that at these early stages the plasmid was a burden on its host; fitness was imparted later on in plant maturation (Lilley and Bailey, 1997b). Therefore, IVET studies are only part of the tool set needed to identify the plasmid loci responsible for the increased host fitness in the natural environment. Obviously, the ideal situation would be to assess transcription and protein expression *in planta* using transcriptomics and proteomic techniques. Unfortunately, the major limitation in such approaches is the recovery of enough cells to make this possible (see discussion, chapter 4). Due to the limitations of the aforementioned studies alternative approaches need to be sought to investigate the ecological relevance of these plasmids.

One such approach would be to sequence more pQBR plasmids. Candidates for sequencing would include representatives from the other two dominant plasmid groups, that have been isolated other successive seasons (group III and IV) (Lilley et al., 1996). Here representatives would be pQBR55 (group III) and pQBR57 (group IV) that were investigated in this thesis. It is also suggested that more group I plasmids should be sequenced. Based on the findings of this thesis, it is proposed these should, at least, include pQBR44 and pQBR47. A practical limitation to sequencing more pQBR plasmids is the bottleneck caused by difficulties of extracting and purifying the plasmid from host DNA, as well as obtaining a reliable sequencing library. This has been demonstrated by in house experience as well as the authors own experiences, whereby numerous successful plasmid extractions have failed to yield genomic sequences (data not shown). However, if the four plasmids proposed could be sequenced, the immediate benefits provided by *in silico* analysis, would include: Firstly, a clearer understating of the genetic structure of the pQBR group I plasmids, as well as the mechanism driving the structural instability of this group. Secondly, confirmation that the group I have an independent evolutionary history from groups III and IV, as well as assessing the genetic distinctness between the group III and IV themselves. Thirdly, determination of whether non-group I plasmids are equally as enigmatic regarding their encoded ecological traits as pQBR103.

The longer-term benefit of obtaining more plasmid genomes is that it provides an opportunity to investigate ecologically relevant plasmid gene/loci *in situ*. At the Oxford field site, identifying members of the dominant pQBR groups over successive years makes it plausible that the pQBR plasmid community is stable, and still exists at the site. In this PhD study, with the anticipation of further plasmid genomes being sequenced, a sugar beet field study was initiated at the Oxford site. This was designed to determine whether genes/loci of ecological relevance to the phytosphere host population, could be identified based on following the fate of the genes themselves *in planta*, both in space (e.g. rhizosphere, rhizoplane) and time (the growing season). However, with only one representative from a single group, whose appearance in the phytopsphere population is correlated to a single plant growth stage (Lilley and Bailey, 1997b), such a study is limited. In comparison, with the sequencing of at least the representatives from the different dominant groups (which display differing temporal successions), such a study is likely to be more insightful in investigating the ecology of these plasmids in the phytosphere.

Given time and funding a multi-plasmid array, which represents all the CDSs predicted for each of the sequenced plasmid genomes, could be constructed and employed to investigate gene frequency *in planta*. This experimental design assumes; Firstly that there is high degree of plasmid mixing and rearrangements (at least within group if not between) which has been demonstrated for the group I (Chapter 3); Secondly, minimal sequence variation between shared loci within groups, which has been shown for the group I (chapter 2) and previously for the group III (Turner et al., 2002); Thirdly, that plasmid copy number is equal (which is testable as part of the experimental design). If these assumptions hold true it may be possible to pool endogenously isolated plasmids from the culturable phytosphere proportion (e.g. pool 5 isolated plasmids from the pseudomonads) for each sample taken, per temporal and spatial point. The constructed array could then be interrogated with each pool "quantitatively" using a two channel design (control plus pooled sample), not to detect gene expression, but to detect gene copy number in each sample. Although such an approach is ambitious given appropriate controls and extensive statistical testing it may at least be possible. The benefit of the characterised collection of pQBR plasmids is the feasibility of such an approach can be tested. If this approach was determined feasible, and if sufficient samples were taken per spatial/temporal point (e.g. 50, which would represent 250 plasmids per point), it may be possible to correlate gene frequency within space and time. The rationale for doing so being that genes/loci of host benefit at that spatial/temporal point will be at a relatively higher frequency than those that are not. Based on these finding in may be possible to further discuss the potential ecological role of these genes, based on correlating frequency data with the understanding of plant physiology i.e. what gene/loci are more prevalent following the release of plant hormones for leaf abscission?

In addition to assessing genes of ecological benefit, a population genetics study would also allow a better appreciation of the seasonality and genetic structure of these different pQBR groups. In this thesis it has been shown that pQBR103 likely shares a largely independent evolutionary history from the group III and IV plasmids. If this is correct it would be possible to follow fate of the different *rep* types *in planta*. This would enable the overlap and ultimate seasonality demonstrated for these groups, to be assessed to a higher resolution than has been afforded previously. Further to this, by correlating genes/loci to these *rep* types it will be able to identify the genetic "core" of these different groups. The results from this thesis strongly point to this approach of studying plasmid population genetics at scale (in terms of number of plasmids assayed) *in planta*. Doing so for such a large group of ecologically well studied plasmids will not only be insightful, but also unprecedented.

References

Abeles, A. L., Friedman, S. A. & Austin, S. J. (1985). Partition of unit-copy miniplasmids to daughter cells .III. The DNA sequence and functional organization of the P1 partition region. *J Mol Biol* **185**, 261-272.

Achtman, M., Kennedy, N. & Skurray, R. (1977). Cell-cell interactions in conjugating *Escherichia coli*: role of traT Protein in Surface Exclusion. *Proc Natl Acad Sci U S A* **74**, 5104-5108.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.

Ashcroft, A. E. (2003). Protein and peptide identification: the role of mass spectrometry in proteomics. *Nat Prod Rep* **20**, 202-215.

Bailey, M. J., Lilley, A. K., Thompson, I. P., Rainey, P. B. & Ellis, R. J. (1995). Site directed chromosomal marking of a fluorescent pseudomonad isolated from the phytosphere of sugar beet; Stability and potential for marker gene transfer. *Mol Ecol* **4**, 755-763.

Bailey, M., Lilley, A. & Diaper, J. (1996). Gene transfer between micro-organisms in the phyllosphere. In *Aerial plant surface microbiology*, pp. 103-123. Edited by C. Morris, C. Nguyen-The & P. Nicot. New York, N.Y: Plenum Press.

Bailey, M. J., Rainey, P. B., Zhang, X. X. & Lilley, A. K. (2001). Population dynamics, gene transfer and gene expression in plasmids, the role of the horizontal gene pool in local adaptation at the plant surface. In *Phytosphere Microbiology*, pp. 173-191. Edited by Lindow S, Hecht-Poiner E and Elliot, V. St Paul M.N. USA: American Phytopathological Society.

Bale, M. J., Fry, J. C. & Day, M. J. (1987). Plasmid transfer between strains of *Pseudomonas aeruginosa* on membrane filters attached to river stones. *J Gen Microbiol* **133**, 3099-3107.

Bale, M. J., Fry, J. C. & Day, M. J. (1988). Transfer and occurrence of large mercury resistance plasmids in river epilithon. *Appl Environ Microbiol* **54**, 972-978.

Barkay, T., Miller, S. M. & Summers, A. O. (2003). Bacterial mercury resistance from atoms to ecosystems. *Fems Microbiol Rev* **27**, 355-384.

Beattie, G. A. & Lindow, S. E. (1995). The secret life of foliar bacterial pathogens on leaves. *Annu Rev Phytopathol* **33**, 145-172.

Bender, C. L. & Cooksey, D. A. (1986). Indigenous plasmids in *Pseudomonas syringae* pv tomato: conjugative transfer and role in copper resistance. *J Bacteriol* **165**, 534-541.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. (2006). GenBank. *Nucleic Acids Res* **34**, D16-20.

Bentley, W. E., Mirjalili, N., Andersen, D. C., Davis, R. H. & Kompala, D. S. (1990). Plasmid-encoded protein: the principal factor in the "metabolic burden" associated with recombinant Bacteria. *Biotechnol Bioeng* **35**, 668-681.

Berg, O. G. & Kurland, C. G. (2002). Evolution of microbial genomes: sequence acquisition and loss. *Mol Biol Evol* **19**, 2265-2276.

Berggren, K., Chernokalskaya, E., Steinberg, T. H., Kemper, C., Lopez, M. F., Diwu, Z., Haugland, R. P. & Patton, W. F. (2000). Background-free, high sensitivity staining of proteins in one- and two-dimensional sodium dodecyl sulfate-polyacrylamide-gels using a luminescent ruthenium complex. *Electrophoresis* **21**, 2509-2521.

Bergstrom, C. T., Lipsitch, M. & Levin, B. R. (2000). Natural selection, infectious transfer and the existence conditions for bacterial plasmids. *Genetics* **155**, 1505-1519.

Bignell, C. & Thomas, C. M. (2001). The bacterial ParA-ParB partitioning proteins. *J Biotechnol* **91**, 1-34.

Bjorklof, K., Suoniemi, A., Haahtela, K. & Romantschuk, M. (1995). High-frequency of conjugation versus plasmid segregation of RP1 in epiphytic *Pseudomonas syringae* populations. *Microbiology (Reading, Engl)* **141**, 2719-2727.

Bonfield, J. K., Smith, K. F. & Staden, R. (1995). A new DNA sequence assembly program. *Nucleic Acids Res* **23**, 4992-4999.

Bourzac, K. M. & Guillemin, K. (2005). *Helicobacter pylori* host cell interactions mediated by type IV secretion. *Cell Microbiol* **7**, 911-919.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72**, 248-254.

Bradley, D. E., Taylor, D. E. & Cohen, D. R. (1980). Specification of surface mating systems among conjugative drug resistance plasmids in *Escherichia coli* K-12. *J Bacteriol* **143**, 1466-1470.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. & Vingron, M. (2001). Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat Genet* **29**, 365-371.

Brown, T. (1994-2005). Dot blot hybridisations. In *Current Protocols in Molecular Biology*. Edited by F. Ausubel. New york, N.Y., USA: Jonh Whiley and Sons, inc.

Buell, C. R., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I. T., Gwinn, M. L., Dodson, R. J., Deboy, R. T., Durkin, A. S., Kolonay, J. F., Madupu, R., Daugherty, S., Brinkac, L., Beanan, M. J., Haft, D. H., Nelson, W. C., Davidsen, T., Zafar, N.,

Zhou, L., Liu, J., Yuan, Q., Khouri, H., Fedorova, N., Tran, B., Russell, D., Berry, K., Utterback, T., Van Aken, S. E., Feldblyum, T. V., D'Ascenzo, M., Deng, W. L., Ramos, A. R., Alfano, J. R., Cartinhour, S., Chatterjee, A. K., Delaney, T. P., Lazarowitz, S. G., Martin, G. B., Schneider, D. J., Tang, X., Bender, C. L., White, O., Fraser, C. M. & Collmer, A. (2003). The complete genome sequence of the Arabidopsis and tomato pathogen *Pseudomonas syringae* pv. tomato DC3000. *Proc Natl Acad Sci U S A* **100**, 10181-10186.

Byrd, D. R. & Matson, S. W. (1997). Nicking by transesterification: the reaction catalysed by a relaxase. *Mol Microbiol* **25**, 1011-1022.

Cabezon, E., Sastre, J. I. & de laCruz, F. (1997). Genetic evidence of a coupling role for the TraG protein family in bacterial conjugation. *Mol Gen Genet* **254**, 400-406.

Call, D. R., Kang, M. S., Daniels, J. & Besser, T. E. (2006). Assessing genetic diversity in plasmids from *Escherichia coli* and *Salmonella enterica* using a mixed-plasmid microarray. *J Appl Microbiol* **100**, 15-28.

Casjens, S., Palmer, N., van Vugt, R., Huang, W. M., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., Dodson, R. J., Haft, D., Hickey, E., Gwinn, M., White, O. & Fraser, C. M. (2000). A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi. Mol Microbiol* **35**, 490-516.

Chen, H. C., Higgins, J., Oresnik, I. J., Hynes, M. F., Natera, S., Djordjevic, M. A., Weinman, J. J. & Rolfe, B. G. (2000). Proteome analysis demonstrates complex replicon and luteolin interactions in pSyma-cured derivatives of *Sinorhizobium meliloti* strain 2011. *Electrophoresis* **21**, 3833-3842.

Chen, Y. T., Chang, H. Y., Lu, C. L. & Peng, H. L. (2004). Evolutionary analysis of the two-component systems in *Pseudomonas aeruginosa* PAO1. *J Mol Evol* **59**, 725-737.

Christie, P. J. & Vogel, J. P. (2000). Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. *Trends Microbiol* **8**, 354-360.

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R. M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M. A., Rajandream, M. A., Rutherford, K. M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J. R. & Barrell, B. G. (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007-1011.

Couturier, M., Bex, F., Bergquist, P. L. & Maas, W. K. (1988). Identification and Classification of Bacterial Plasmids. *Microbiol Rev* **52**, 375-395.

Dahlberg, C., Linberg, C., Torsvik, V. L. & Hermansson, M. (1997). Conjugative plasmids isolated from bacteria in marine environments show various degrees of homology to each other and are not closely related to well-characterized plasmids. *Appl Environ Microbiol* **63**, 4692-4697.

Dahlberg, C. & Chao, L. (2003). Amelioration of the cost of conjugative plasmid carriage in *Eschericha coli* K12. *Genetics* **165**, 1641-1649.

Davis, M. A., Martin, K. A. & Austin, S. J. (1992). Biochemical activities of the ParA Partition protein of the P1 plasmid. *Mol Microbiol* **6**, 1141-1147.

Davison, J. (1999). Genetic exchange between bacteria in the environment. *Plasmid* **42**, 73-91.

de Weert, S., Vermeiren, H., Mulders, I. H. M., Kuiper, I., Hendrickx, N., Bloemberg, G. V., Vanderleyden, J., De Mot, R. & Lugtenberg, B. J. J. (2002). Flagella-driven chemotaxis towards exudate components is an important trait for tomato root colonization by *Pseudomonas fluorescens*. *Mol Plant Microbe Interact* **15**, 1173-1180.

Dekio, I., Hayashi, H., Sakamoto, M., Kitahara, M., Nishikawa, T., Suematsu, M. & Benno, Y. (2005). Detection of potentially novel bacterial components of the human skin microbiota using culture-independent molecular profiling. *J Med Microbiol* **54**, 1231-1238.

Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636-4641.

Dennis, J. J. & Zylstra, G. J. (2004). Complete sequence and genetic organization of pDTG1, the 83 kilobase naphthalene degradation plasmid from *Pseudomonas putida* strain NCIB 9816-4. *J Mol Biol* **341**, 753-768.

Dennis, J. J. (2005). The evolution of IncP catabolic plasmids. *Curr Opin Biotechnol* **16**, 291-298.

Doolittle, R. F. (2002). Biodiversity: microbial genomes multiply. *Nature* **416**, 697-700.

Dorrell, N., Mangan, J. A., Laing, K. G., Hinds, J., Linton, D., Al-Ghusein, H., Barrell, B. G., Parkhill, J., Stoker, N. G., Karlyshev, A. V., Butcher, P. D. & Wren, B. W. (2001). Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res* **11**, 1706-1715.

Eberhard, W. G. (1990). Evolution in bacterial plasmids and levels of selection. *Q Rev Biol* **65**, 3-22.

Edwards, R. A. & Rohwer, F. (2005). Viral metagenomics. *Nat Rev Microbiol* **3**, 504-510.

Enright, A. J., Van Dongen, S. & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584.

Espinosa, M., Cohen, M., Couturier, M., del Solar, G., Diaz-Orejas, R., Giraldo, R., Janniere, L., Miller, C., Osborn, M. & Thomas, C. M. (2000). Plasmid Replication and Copy Number Control. In *The Horizontal Gene Pool: Bacterial plasmids and gene spread*, pp. 87-174. Edited by C. M. Thomas. Amsterdam: Harward academic publishers.

Feil, E. J. & Spratt, B. G. (2001). Recombination and the population structures of bacterial pathogens. *Annu Rev Microbiol* **55**, 561-590.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M., Ashburner, M., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., dePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Oliver Glöckner, F., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kyripides, N., Leebens-Mack, J., Lewis, S., Liste, A., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Moxon, R., Nelson, K., Parkhill, J., Sansone, S., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., Wilson, G. & Wipat, A. Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification. *Submitted.*

Field, D. & Hughes, J. (2005). Cataloguing our current genome collection. *Microbiology (Reading, Engl)* **151**, 1016-1019.

Finan, T. M., Weidner, S., Wong, K., Buhrmester, J., Chain, P., Vorholter, F. J., Hernandez-Lucas, I., Becker, A., Cowie, A., Gouzy, J., Golding, B. & Puhler, A. (2001). The complete sequence of the 1,683-kb pSymB megaplasmid from the N-2-fixing endosymbiont *Sinorhizobium meliloti*. *Proc Natl Acad Sci U S A* **98**, 9889-9894.

Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. & Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, D247-251.

Fitzgerald, J. R., Sturdevant, D. E., Mackie, S. M., Gill, S. R. & Musser, J. M. (2001). Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc Natl Acad Sci U S A* **98**, 8821-8826.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C.

M., Smith, H. O. & Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

Franklin, F. C. H., Bagdasarian, M., Bagdasarian, M. M. & Timmis, K. N. (1981). Molecular and functional analysis of the TOL plasmid pWWO from *Pseudomonas putida* and cloning of genes for the entire regulated aromatic ring meta-cleavage pathway. *Proc Natl Acad Sci U S A* **78**, 7458-7462.

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., Peterson, J., Kerlavage, A. R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M. D., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fuji, C., Cotton, M. D., Horst, K., Roberts, K., Hatch, B., Smith, H. O. & Venter, J. C. (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580-586.

Fraser-Liggett, C. M. (2005). Insights on biology and evolution from microbial genome sequencing. *Genome Res* **15**, 1603-1610.

Freiberg, C., Fellay, R., Bairoch, A., Broughton, W. J., Rosenthal, A. & Perret, X. (1997). Molecular basis of symbiosis between *Rhizobium* and legumes. *Nature* **387**, 394-401.

Frost, L. S., Ippenihler, K. & Skurray, R. A. (1994). Analysis of the sequence and gene products of the transfer region of the F-sex factor. *Microbiol Rev* **58**, 162-210.

Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* **3**, 722-732.

Fry, J. C. & Day, M. J. (1990). Plasmid trasfer in the epilithon. In *Bacterial Genetics in Natural Environments*, pp. 55-80. Edited by J. C. Fry & M. J. Day. London, UK: Chapman and Hall.

Fry, J. (2000). Bacterial diversity and 'unculturables'. *SGM Microbiology today* **27**, 186-189.

Galibert, F., Finan, T. M., Long, S. R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M. J., Becker, A., Boistard, P., Bothe, G., Boutry, M., Bowser, L., Buhrmester, J., Cadieu, E., Capela, D., Chain, P., Cowie, A., Davis, R. W., Dreano, S., Federspiel, N. A., Fisher, R. F., Gloux, S., Godrie, T., Goffeau, A., Golding, B., Gouzy, J., Gurjal, M., Hernandez-Lucas, I., Hong, A., Huizar, L., Hyman, R. W., Jones, T., Kahn, D., Kahn, M. L., Kalman, S., Keating, D. H., Kiss, E., Komp, C., Lalaure, V., Masuy, D., Palm, C., Peck, M. C., Pohl, T. M., Portetelle, D., Purnelle, B., Ramsperger, U., Surzycki, R., Thebault, P., Vandenbol, M., Vorholter, F. J., Weidner, S., Wells, D. H., Wong, K., Yeh, K. C. & Batut, J. (2001). The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293**, 668-672.

Garrity, G. (2001). Bergey's Manual of Systematic Bacteriology. New York, N.Y. USA: Springer.

Gilmour, M. W., Thomson, N. R., Sanders, M., Parkhill, J. & Taylor, D. E. (2004). The complete nucleotide sequence of the resistance plasmid R478: defining the backbone components of incompatibility group H conjugative plasmids through comparative genomics. *Plasmid* **52**, 182-202.

Giuntini, E., Mengoni, A., De Filippo, C., Cavalieri, D., Aubin-Horth, N., Landry, C. R., Becker, A. & Bazzicalupo, M. (2005). Large-scale genetic variation of the symbiosis-required megaplasmid pSymA revealed by comparative genomic analysis of *Sinorhizobium meliloti* natural strains. *BMC Genomics* **6**, 158.

Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**, 2226-2238.

Gogarten, J. P. & Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**, 679-687.

Gomis-Ruth, F. X., de la Cruz, F. & Coll, M. (2002). Structure and role of coupling proteins in conjugal DNA transfer. *Res Microbiol* **153**, 199-204.

Gonzalez, V., Bustos, P., Ramirez-Romero, M. A., Medrano-Soto, A., Salgado, H., Hernandez-Gonzalez, I., Hernandez-Celis, J. C., Quintero, V., Moreno-Hagelsieb, G., Girard, L., Rodriguez, O., Flores, M., Cevallos, M. A., Collado-Vides, J., Romero, D. & Davila, G. (2003). The mosaic structure of the symbiotic plasmid of *Rhizobium etli* CFN42 and its relation to other symbiotic genome compartments. *Genome Biol* **4(6)**;R36.

Gonzalez, V., Santamaria, R. I., Bustos, P., Hernandez-Gonzalez, I., Medrano-Soto, A., Moreno-Hagelsieb, G., Janga, S. C., Ramirez, M. A., Jimenez-Jacinto, V., Collado-Vides, J. & Davila, G. (2006). The partitioned *Rhizobium etli* genome: Genetic and metabolic redundancy in seven interacting replicons. *Proc Natl Acad Sci U S A* **103**, 3834-3839.

Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B. S., Cao, Y., Askenazi, M., Halling, C., Mullin, L., Houmiel, K., Gordon, J., Vaudin, M., Iartchouk, O., Epp, A., Liu, F., Wollam, C., Allinger, M., Doughty, D., Scott, C., Lappas, C., Markelz, B., Flanagan, C., Crowell, C., Gurson, J., Lomo, C., Sear, C., Strub, G., Cielo, C. & Slater, S. (2001). Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294**, 2323-2328.

Greated, A., Lambertsen, L., Williams, P. A. & Thomas, C. M. (2002). Complete sequence of the IncP-9 TOL plasmid pWW0 from *Pseudomonas putida. Environ Microbiol* **4**, 856-871.

Griffith, F. (1928). The significance of pneumococcal types. *J Hyg(Lond)* **27**, 113-159.

Griffiths, R. I., Whiteley, A. S., O'Donnell, A. G. & Bailey, M. J. (2000). Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Appl Environ Microbiol* **66**, 5488-5491.

Grimm, A. C. & Harwood, C. S. (1999). NahY, a catabolic plasmid-encoded receptor required for chemotaxis of *Pseudomonas putida* to the aromatic hydrocarbon naphthalene. *J Bacteriol* **181**, 3310-3316.

Guerreiro, N., Redmond, J. W., Rolfe, B. G. & Djordjevic, M. A. (1997). New *Rhizobium leguminosarum* flavonoid-induced proteins revealed by proteome analysis of differentially displayed proteins. *Mol Plant Microbe Interact* **10**, 506-516.

Guerreiro, N., Stepkowski, T., Rolfe, B. G. & Djordjevic, M. A. (1998). Determination of plasmid-encoded functions in *Rhizobium leguminosarum* biovar trifolii using proteome analysis of plasmid-cured derivatives. *Electrophoresis* **19**, 1972-1979.

Guerreiro, N., Djordjevic, M. A. & Rolfe, B. G. (1999). Proteome analysis of the model microsymbiont *Sinorhizobium meliloti*: isolation and characterisation of novel proteins. *Electrophoresis* **20**, 818-825.

Gutierrez, P., Li, Y., Osborne, M. J., Pomerantseva, E., Liu, Q. & Gehring, K. (2005). Solution structure of the carbon storage regulator protein CsrA from *Escherichia coli. J Bacteriol* **187**, 3496-3501.

Harayama, S. (1994). Codon usage patterns suggest independent evolution of two catabolic operons on toluene degradative Plasmid TOL pWW0 of *Pseudomonas putida. J Mol Evol* **38**, 328-335.

Hayes, F. (2000). The partition system of multidrug resistance plasmid TP228 includes a novel protein that epitomizes an evolutionarily distinct subgroup of the ParA superfamily. *Mol Microbiol* **37**, 528-541.

Helgason, E., Okstad, O. A., Caugant, D. A., Johansen, H. A., Fouet, A., Mock, M., Hegna, I. & Kolsto (2000). *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*: one species on the basis of genetic evidence. *Appl Environ Microbiol* **66**, 2627-2630.

Hendson, M., Purcell, A. H., Chen, D. Q., Smart, C., Guilhabert, M. & Kirkpatrick, B. (2001). Genetic diversity of Pierce's disease strains and other pathotypes of *Xylella fastidiosa. Appl Environ Microbiol* **67**, 895-903.

Hill, K. E., Weightman, A. J. & Fry, J. C. (1992). Isolation and screening of plasmids from the epilithon which mobilize recombinant plasmid pD10. *Appl Environ Microbiol* **58**, 1292-1300.

Hoffmaster, A. R., Ravel, J., Rasko, D. A., Chapman, G. D., Chute, M. D., Marston, C. K., De, B. K., Sacchi, C. T., Fitzgerald, C., Mayer, L. W., Maiden, M. C., Priest, F. G., Barker, M., Jiang, L., Cer, R. Z., Rilstone, J., Peterson, S. N.,

Weyant, R. S., Galloway, D. R., Read, T. D., Popovic, T. & Fraser, C. M. (2004). Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. *Proc Natl Acad Sci U S A* **101**, 8449-8454.

Honore, P., Granjeaud, S., Tagett, R., Deraco, S., Beaudoing, E., Rougemont, J., Debono, S. & Hingamp, P. (2006). MicroArray Facility: a laboratory information management system with extended support for Nylon based technologies. *BMC Genomics* 7, 240.

Houlden, A. (2005).*Bacterial and fungal diversity effects and the activity of biocontrol agents in the rhizosphere of crop plants.* pp. 243. PhD thesis: Cardiff University.

Hutchison, C. A. & Venter, J. C. (2006). Single-cell genomics. *Nat Biotechnol* **24**, 657-658.

Janssen, P. H. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl Environ Microbiol* **72**, 1719-1728.

Jiang, S. C. & Paul, J. H. (1998). Gene transfer by transduction in the marine environment. *Appl Environ Microbiol* **64**, 2780-2787.

Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A., Idesawa, K., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Mochizuki, Y., Nakayama, S., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada, M. & Tabata, S. (2000). Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti. DNA Res* **7**, 331-338.

Karnholz, A., Hoefler, C., Odenbreit, S., Fischer, W., Hofreuter, D. & Haas, R. (2006). Functional and topological characterization of novel components of the comB DNA transformation competence system in *Helicobacter pylori. J Bacteriol* **188**, 882-893.

Khesin, R. B. & Karasyova, E. V. (1984). Mercury resistant plasmids in Bacteria from a mercury and antimony deposit area. *Mol Gen Genet* **197**, 280-285.

Kim, C. C., Joyce, E. A., Chan, K. & Falkow, S. (2002). Improved analytical methods for microarray-based genome-composition analysis. *Genome Biol* 3(11); R65.

Kobayashi, N. & Bailey, M. J. (1994). Plasmids isolated from the sugar-beet phyllosphere show little or no homology to molecular probes currently available for plasmid typing. *Microbiology(Reading,Engl)* **140**, 289-296.

Konstantinidis, K. T. & Tiedje, J. M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci U S A* **101**, 3160-3165.

Krakauer, D. C. & Plotkin, J. B. (2002). Redundancy, antiredundancy, and the robustness of genomes. *Proc Natl Acad Sci U S A* **99**, 1405-1409.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J. & Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* **29**, 4633-4642.

Kyrpides, N. C. (1999). Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics* **15**, 773-774.

Lai, E. M., Shih, H. W., Wen, S. R., Cheng, M. W., Hwang, H. H. & Chiu, S. H. (2006). Proteomic analysis of *Agrobacterium tumefaciens* response to the Vir gene inducer acetosyringone. *Proteomics* **6**, 4130-4136.

Lanfermeijer, F. C., Vanoene, M. A. & Borstlap, A. C. (1992). Compartmental analysis of amino-acid release from attached and detached pea seed coats. *Planta* **187**, 75-82.

Langlois, P., Bourassa, S., Poirier, G. G. & Beaulieu, C. (2003). Identification of *Streptomyces coelicolor* proteins that are differentially expressed in the presence of plant material. *Appl Environ Microbiol* **69**, 1884-1889.

Lawley, T. D., Klimke, W. A., Gubbins, M. J. & Frost, L. S. (2003)a. F factor conjugation is a true type IV secretion system. *FEMS Microbiol Lett* **224**, 1-15.

Lawley, T. D., Gilmour, M. W., Gunton, J. E., Tracz, D. M. & Taylor, D. E. (2003)b. Functional and mutational analysis of conjugative transfer region 2 (Tra2) from the IncHI1 plasmid R27. *J Bacteriol* **185**, 581-591.

Lawrence, J. G. & Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**, 383-397.

Lawrence, J. G. & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A* **95**, 9413-9417.

Lederberg, J. & Tatum, E. (1946). Gene recombination in *E. coli*. *Nature* **158**, 558.

Lederberg, J., Lederberg, E. M., Zinder, N. D. & Lively, E. R. (1951). Recombination analysis of bacterial heredity. *Cold Spring Harb Symp Quant Biol* **16**, 413-443.

Lengeler, W. J., Drews, D. & Schlegal, H. G. (1999). *Biology of the Prokaryotes.* Oxford UK: Blackwell Publishing ltd.

Lenski, R. E. & Bouma, J. E. (1987). Effects of segregation and selection on instability of plasmid pACYC184 *Escherichia coli* B. *J Bacteriol* **169**, 5314-5316.

Leplae, R., Hebrant, A., Wodak, S. J. & Toussaint, A. (2004). ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res* **32**, D45-49.

Leplae, R., Lima-Mendez, G. & Toussaint, A. (2006). A first global analysis of plasmid encoded proteins in the ACLAME database. *FEMS Microbiol Rev* **30**, 980-994.

Levin, B. R., Stewart, F. M. & Rice, V. A. (1979). Kinetics of conjugative plasmid transmission: fit of a simple mass action model. *Plasmid* **2,** 247-260.

Levin, B. R. & Bergstrom, C. T. (2000). Bacteria are different: Observations, interpretations, speculations, and opinions about the mechanisms of adaptive evolution in prokaryotes. *Proc Natl Acad Sci U S A* **97**, 6981-6985.

Lilley, A. K., Fry, J. C., Day, M. J. & Bailey, M. J. (1994). *In situ* transfer of an exogenously isolated plasmid between *Pseudomonas Spp* in sugar beet rhizosphere. *Microbiology(Reading, Engl)* **140**, 27-33.

Lilley, A. K., Bailey, M. J., Day, M. J. & Fry, J. C. (1996). Diversity of mercury resistance plasmids obtained by exogenous isolation from the bacteria of sugar beet in three successive years. *FEMS Microbiol Ecol* **20**, 211-227.

Lilley, A. K. & Bailey, M. J. (1997)a. The acquisition of indigenous plasmids by a genetically marked pseudomonad population colonizing the sugar beet phytosphere is related to local environmental conditions. *Appl Environ Microbiol* **63**, 1577-1583.

Lilley, A. K. & Bailey, M. J. (1997)b. Impact of plasmid pQBR103 acquisition and carriage on the phytosphere fitness of *Pseudomonas fluorescens* SBW25: Burden and benefit. *Appl Environ Microbiol* **63**, 1584-1587.

Lindow, S. E. & Brandl, M. T. (2003). Microbiology of the phyllosphere. *Appl Environ Microbiol* **69**, 1875-1883.

Liolios, K., Tavernarakis, N., Hugenholtz, P. & Kyrpides, N. C. (2006). The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res* **34**, D332-334.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nat Genet* **21**, 20-24.

Llanes, C., Gabant, P., Couturier, M., Bayer, L. & Plesiat, P. (1996). Molecular analysis of the replication elements of the broad-host range RepA/C replicon. *Plasmid* **36**, 26-35.

Lobocka, M. B., Rose, D. J., Plunkett, G., Rusin, M., Samojedny, A., Lehnherr, H., Yarmolinsky, M. B. & Blattner, F. R. (2004). Genome of bacteriophage P1. *J Bacteriol* **186**, 7032-7068.

Lorenz, M. G. & Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol Rev* **58**, 563-602.

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964.

Lucchini, S., Thompson, A. & Hinton, J. C. (2001). Microarrays for microbiologists. *Microbiology (Reading, Engl)* **147**, 1403-1414.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Forster, W., Brettske, I., Gerber, S., Ginhart, A. W., Gross, O., Grumann, S., Hermann, S., Jost, R., Konig, A., Liss, T., Lussmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A. & Schleifer, K. H. (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**, 1363-1371.

Lux, R. & Shi, W. Y. (2004). Chemotaxis-guided movements in bacteria. *Crit Rev Oral Biol Med* **15**, 207-220.

Mahan, M. J., Slauch, J. M. & Mekalanos, J. J. (1993). Selection of Bacterial virulence genes that are specifically induced in host tissues. *Science* **259**, 686-688.

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982). *Molecular cloning: a laboratory manual*. New York, N.Y., USA. Cold Spring Harbour Laboratory.

Marx, C. J., Miller, J. A., Chistoserdova, L. & Lidstrom, M. E. (2004). Multiple formaldehyde oxidation/detoxification pathways in *Burkholderia fungorum* LB400. *J Bacteriol* **186**, 2173-2178.

McAllister, C. F. & Stephens, D. S. (1993). Analysis in *Neisseria meningitidis* and other *Neisseria* species of genes homologous to the FKBP immunophilin family. *Mol Microbiol* **10**, 13-23.

Mcclure, N. C., Weightman, A. J. & Fry, J. C. (1989). Survival of *Pseudomonas putida* UWC1 containing cloned catabolic genes in a model activated-sludge unit. *Appl Environ Microbiol* **55**, 2627-2634.

Mergeay, M., Lejeune, P., Sadouk, A., Gerits, J. & Fabry, L. (1987). Shuttle transfer (or Retrotransfer) of chromosomal markers mediated by plasmid pULB113. *Mol Gen Genet* **209**, 61-70.

Mindlin, S., Minakhin, L., Petrova, M., Kholodii, G., Minakhina, S., Gorlenko, Z. & Nikiforov, V. (2005). Present-day mercury resistance transposons are common in Bacteria preserved in permafrost grounds since the Upper Pleistocene. *Res Microbiol* **156**, 994-1004.

Modi, R. I. & Adams, J. (1991). Coevolution in bacterial-plasmid populations. *Evolution* **45**, 656-667.

Molbak, L., Tett, A., Ussery, D. W., Wall, K., Turner, S., Bailey, M. & Field, D. (2003). The plasmid genome database. *Microbiology (Reading,Engl)* **149**, 3043-3045.

Monteiro-Vitorello, C. B., De Oliveira, M. C., Zerillo, M. M., Varani, A. M., Civerolo, E. & Van Sluys, M. A. (2005). *Xylella* and *Xanthomonas* Mobil'omics. *OMICS* **9**, 146-159.

Morgan, J. A. W., Bending, G. D. & White, P. J. (2005). Biological costs and benefits to plant-microbe interactions in the rhizosphere. *J Exp Bot* **56**, 1729-1739.

Mori, H., Kondo, A., Ohshima, A., Ogura, T. & Hiraga, S. (1986). Structure and function of the F-plasmid genes essential for partitioning. *J Mol Biol* **192**, 1-15.

Muller, H. J. (1932). Some genetic aspects of sex. *Am Nat* **66**, 118-138.

Nelson, K. E., Weinel, C., Paulsen, I. T., Dodson, R. J., Hilbert, H., Martins dos Santos, V. A., Fouts, D. E., Gill, S. R., Pop, M., Holmes, M., Brinkac, L., Beanan, M., DeBoy, R. T., Daugherty, S., Kolonay, J., Madupu, R., Nelson, W., White, O., Peterson, J., Khouri, H., Hance, I., Chris Lee, P., Holtzapple, E., Scanlan, D., Tran, K., Moazzez, A., Utterback, T., Rizzo, M., Lee, K., Kosack, D., Moestl, D., Wedler, H., Lauber, J., Stjepandic, D., Hoheisel, J., Straetz, M., Heim, S., Kiewitz, C., Eisen, J. A., Timmis, K. N., Dusterhoft, A., Tummler, B. & Fraser, C. M. (2002). Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol* **4**, 799-808.

Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250**, 4007-4021.

Osborn, A. M., Bruce, K. D., Strike, P. & Ritchie, D. A. (1997). Distribution, diversity and evolution of the bacterial mercury resistance (mer) operon. *FEMS Microbiol Rev* **19**, 239-262.

Osborn, A. M., Tatley, F. M. D., Steyn, L. M., Pickup, R. W. & Saunders, J. R. (2000). Mosaic plasmids and mosaic replicons: evolutionary lessons from the analysis of genetic diversity in IncFII-related replicons. *Microbiology (Reading, Engl)* **146**, 2267-2275.

Osbourn, A. E., Barber, C. E. & Daniels, M. J. (1987). Identification of plant Induced genes of the bacterial pathogen *Xanthomonas campestris* pathovar campestris using a promoter-probe plasmid. *EMBO J* **6**, 23-28.

Ow, D. S. W., Nissom, P. M., Philp, R., Oh, S. K. W. & Yap, M. G. S. (2006). Global transcriptional analysis of metabolic burden due to plasmid maintenance in *Escherichia coli* DH5 alpha during batch fermentation. *Enzyme Microb Technol* **39**, 391-398.

Pandey, G. & Jain, R. K. (2002). Bacterial chemotaxis toward environmental pollutants: Role in bioremediation. *Appl Environ Microbiol* **68**, 5789-5795.

Pandya, S., Iyer, P., Gaitonde, V., Parekh, T. & Desai, A. (1999). Chemotaxis of *Rhizobium* sp.S2 towards *Cajanus cajan* root exudate and its major components. *Curr Microbiol* **38**, 205-209.

Pansegrau, W., Lanka, E., Barth, P. T., Figurski, D. H., Guiney, D. G., Haas, D., Helinski, D. R., Schwab, H., Stanisich, V. A. & Thomas, C. M. (1994). Complete nucleotide sequence of Birmingham IncP alpha plasmids. Compilation and comparative analysis. *J Mol Biol* **239**, 623-663.

Pansegrau, W. & Lanka, E. (1996). Enzymology of DNA transfer by conjugative mechanisms. *Prog Nucleic Acid Res Mol Biol* **54**, 197-251.

Paulsen, I. T., Press, C. M., Ravel, J., Kobayashi, D. Y., Myers, G. S., Mavrodi, D. V., DeBoy, R. T., Seshadri, R., Ren, Q., Madupu, R., Dodson, R. J., Durkin, A. S., Brinkac, L. M., Daugherty, S. C., Sullivan, S. A., Rosovitz, M. J., Gwinn, M. L., Zhou, L., Schneider, D. J., Cartinhour, S. W., Nelson, W. C., Weidman, J., Watkins, K., Tran, K., Khouri, H., Pierson, E. A., Pierson, L. S., 3rd, Thomashow, L. S. & Loper, J. E. (2005). Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nat Biotechnol* **23**, 873-878.

Paulsson, J. & Ehrenberg, M. (1998). Trade-off between segregational stability and metabolic burden: A mathematical model of plasmid ColE1 replication control. *J Mol Biol* **279**, 73-88.

Paulsson, J. (2002). Multileveled selection on plasmid replication. *Genetics* **161**, 1373-1384.

Perna, N. T., Plunkett, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E. J., Davis, N. W., Limk, A., Dimalanta, E. T., Potamousis, K. D., Apodaca, J., Anantharaman, T. S., Lin, J. Y., Yen, G., Schwartz, D. C., Welch, R. A. & Blattner, F. R. (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157 : H7. *Nature* **409**, 529-533.

Pohlman, R. F., Genetti, H. D. & Winans, S. C. (1994). Common ancestry between IncN conjugal transfer genes and macromolecular export systems of plant and animal pathogens. *Mol Microbiol* **14**, 655-668.

Quaiser, A., Ochsenreiter, T., Lanz, C., Schuster, S. C., Treusch, A. H., Eck, J. & Schleper, C. (2003). Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Mol Microbiol* **50**, 563-575.

Rainey, P. B. (1999). Adaptation of *Pseudomonas fluorescens* to the plant rhizosphere. *Environ Microbiol* **1**, 243-257.

Reasoner, D. J. & Geldreich, E. E. (1985). A new medium for the enumeration and subculture of Bacteria from potable water. *Appl Environ Microbiol* **49**, **1**-7.

Rediers, H., Rainey, P. B., Vanderleyden, J. & De Mot, R. (2005). Unraveling the secret lives of Bacteria: Use of *in vivo* expression technology and differential fluorescence induction promoter traps as tools for exploring niche-specific gene expression. *Microbiol Mol Biol Rev* **69 (2)**, 217-61.

Reinikainen, P. & Virkajarvi, I. (1989). *Escherichia coli* growth and plasmid copy numbers in continuous cultivations. *Biotechnol Lett* **11**, 225-230.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277.

Roberts, D. P., Dery, P. D., Yucel, I., Buyer, J., Holtman, M. A. & Kobayashi, D. Y. (1999). Role of pfkA and general carbohydrate catabolism in seed colonization by Enterobacter cloacae. *Appl Environ Microbiol* **65**, 2513-2519.

Roberts, D. P., Dery, P. D., Yucel, I. & Buyer, J. S. (2000). Importance of pfkA for rapid growth of Enterobacter cloacae during colonization of crop seeds. *Appl Environ Microbiol* **66**, 87-91.

Rocha, E. P. C. (2003). An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: From duplications to genome reduction. *Genome Res* **13**, 1123-1132.

Romling, U., Gomelsky, M. & Galperin, M. Y. (2005). C-di-GMP: the dawning of a novel bacterial signalling system. *Mol Microbiol* **57**, 629-639.

Ronchel, M. C., Ramos-Diaz, M. A. & Ramos, J. L. (2000). Retrotransfer of DNA in the rhizosphere. *Environmental Microbiology* **2**, 319-323.

Rozen, S. & Skaletsky, J. (2000). Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, pp. 365-386. Edited by S. Krawetz & S. Misener. Totowa, NJ.

Rozenberg-Arska, M., Salters, E. C., van Strijp, J. A., Hoekstra, W. P. & Verhoef, J. (1984). Degradation of *Escherichia coli* chromosomal and plasmid DNA in serum. *J Gen Microbiol* **130**, 217-222.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944-945.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-425.

Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L. & Falkow, S. (2000). A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc Natl Acad Sci U S A* **97**, 14668-14673.

Schneiker, S., Keller, M., Droge, M., Lanka, E., Puhler, A. & Selbitschka, W. (2001). The genetic organization and evolution of the broad host range mercury resistance plasmid pSB102 isolated from a microbial population residing in the rhizosphere of alfalfa. *Nucleic Acids Res* **29**, 5169-5181.

Schroder, G. & Lanka, E. (2005). The mating pair formation system of conjugative plasmids: A versatile secretion machinery for transfer of proteins and DNA. *Plasmid* **54**, 1-25.

Schuster, P. F., Krabbenhoft, D. P., Naftz, D. L., Cecil, L. D., Olson, M. L., Dewild, J. F., Susong, D. D., Green, J. R. & Abbott, M. L. (2002). Atmospheric mercury deposition during the last 270 years: A glacial ice core record of natural and anthropogenic sources. *Environ Sci Technol* **36**, 2303-2310.

Schwaner, N. E. & Kroer, N. (2001). Effect of plant species on the kinetics of conjugal transfer in the rhizosphere and relation to bacterial metabolic activity. *Microb Ecol* **42**, 458-465.

Segovia, L., Pinero, D., Palacios, R. & Martinezromero, E. (1991). Genetic structure of a soil population of nonsymbiotic *Rhizobium leguminosarum. Appl Environ Microbiol* **57**, 426-433.

Sesma, A., Sundin, G. W. & Murillo, J. (2000). Phylogeny of the replication regions of pPT23A-like plasmids from *Pseudomonas syringae. Microbiology(Reading, Engl)* **146**, 2375-2384.

Sharma, R., Ranjan, R., Kapardar, R. K. & Grover, A. (2005). 'Unculturable' bacterial diversity: An untapped resource. *Curr Sci* **89**, 72-77.

Shaw, L. J. & Burns, R. G. (2005). Rhizodeposition and the enhanced mineralization of 2,4-dichlorophenoxyacetic acid in soil from the *Trifolium pratense* rhizosphere. *Environ Microbiol* **7**, 191-202.

Siew, N. & Fischer, D. (2003). Twenty thousand ORFan microbial protein families for the biologist? *Structure* **11**, 7-9.

Simonsen, L., Gordon, D. M., Stewart, F. M. & Levin, B. R. (1990). Estimating the rate of plasmid transfer: an end-point method. *J Gen Microbiol* **136**, 2319-2325.

Simonsen, L. (1991). The existence conditions for bacterial plasmids: theory and reality. *Microb Ecol* **22**, 187-205.

Singh, B. K., Millard, P., Whiteley, A. S. & Murrell, J. C. (2004). Unravelling rhizosphere-microbial interactions: opportunities and limitations. *Trends Microbiol* **12**, 386-393.

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. & Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet* **17**, 425-428.

Smalla, K., Osborn, A. M., Wellington, E. M. H. (2000). Isolation and characterisation of plasmids from Bacteria. In *The Horizontal Gene Pool: Bacterial plasmids and gene spread*, pp. 87-174. Edited by C. M. Thomas. Amsterdam: Harward academic publishers.

Smets, B. F. & Barkay, T. (2005). Horizontal gene transfer: perspectives at a crossroads of scientific disciplines. *Nat Rev Microbiol* **3**, 675-678.

Smith, J. M., Smith, N. H., Orourke, M. & Spratt, B. G. (1993). How clonal are Bacteria? *Proc Natl Acad Sci U S A* **90**, 4384-4388.

Smoot, J. C., Barbian, K. D., Van Gompel, J. J., Smoot, L. M., Chaussee, M. S., Sylva, G. L., Sturdevant, D. E., Ricklefs, S. M., Porcella, S. F., Parkins, L. D., Beres, S. B., Campbell, D. S., Smith, T. M., Zhang, Q., Kapur, V., Daly, J. A., Veasy, L. G. & Musser, J. M. (2002). Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci U S A* **99**, 4668-4673.

Snyder, L. A. S., Davies, J. K. & Saunders, N. J. (2004). Microarray genomotyping of key experimental strains of *Neisseria gonorrhoeae* reveals gene complement diversity and five new neisserial genes associated with Minimal Mobile Elements. *BMC Genomics* **5**,23

Snyder, L. A. S., Jarvis, S. A. & Saunders, N. J. (2005). Complete and variant forms of the 'gonococcal genetic island' in *Neisseria meningitidis*. *Microbiology(Reading, Engl)* **151**, 4005-4013.

Sobecky, P. A., Mincer, T. J., Chang, M. C. & Helinski, D. R. (1997). Plasmids isolated from marine sediment microbial communities contain replication and incompatibility regions unrelated to those of known plasmid groups. *Appl Environ Microbiol* **63**, 888-895.

Sorensen, S. J., Bailey, M., Hansen, L. H., Kroer, N. & Wuertz, S. (2005). Studying plasmid horizontal transfer in situ: a critical review. *Nat Rev Microbiol* **3**, 700-710.

Sota, M., Yano, H., Ono, A., Miyazaki, R., Ishii, H., Genka, H., Top, E. M. & Tsuda, M. (2006). Genomic and functional analysis of the IncP-9 naphthalene-catabolic plasmid NAH7 and its transposon Tn*4655* suggests catabolic gene spread by a tyrosine recombinase. *J Bacteriol* **188**, 4057-4067.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel-electrophoresis. *J Mol Biol* **98**, 503-&.

Spiers, A. J., Field, D., Bailey, M. & Rainey, P. B. (2001). Notes on designing a partial genomic database: The PfSBW25 Encyclopaedia, a sequence database for *Pseudomonas fluorescens* SBW25. *Microbiology(Reading, Engl)* **147**, 247-249.

Spratt, B. G., Hanage, W. P. & Feil, E. J. (2001). The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr Opin Microbiol* **4**, 602-606.

Staton, J. L. (2002). Homology in character evolution In *Encyclopaedia of life sciences*, vol 9 pp. 196-202, London, UK, Macmillan publishers ltd.

Sterkenburg, A., Prozee, G. A. P., Leegwater, P. A. J. & Wouters, J. T. M. (1984). Expression and loss of the pBR322 Plasmid in *Klebsiella aerogenes* NCTC-418, grown in chemostat culture. *Antonie Van Leeuwenhoek* **50**, 397-404.

Stewart, F. M. & Levin, B. R. (1977). Population biology of bacterial plasmids: apriori conditions for existence of conjugationally transmitted factors. *Genetics* **87**, 209-228.

Summers, D. K., Beton, C. W. & Withers, H. L. (1993). Multicopy plasmid instability: the dimer catastrophe hypothesis. *Mol Microbiol* **8**, 1031-1038.

Summers, D. K. (1996). *The Biology of Plasmids*. Oxford UK: Blackwell publishing ltd.

Summers, D. (1998). Timing, self-control and a sense of direction are the secrets of multicopy plasmid stability. *Mol Microbiol* **29**, 1137-1145.

Sundin, G. W., Demezas, D. H. & Bender, C. L. (1994). Genetic and plasmid diversity within natural populations of *Pseudomonas syringae* with various exposures to copper and streptomycin bactericides. *Appl Environ Microbiol* **60**, 4421-4431.

Sundin, G. W. & Murillo, J. (1999). Functional analysis of the *Pseudomonas syringae* rulAB determinant in tolerance to ultraviolet B (290-320 nm) radiation and distribution of rulAB among *P. syringae* pathovars. *Environ Microbiol* **1**, 75-87.

Suzek, B. E., Ermolaeva, M. D., Schreiber, M. & Salzberg, S. L. (2001). A probabilistic, method for identifying start codons in bacterial genomes. *Bioinformatics* **17**, 1123-1130.

Szurmant, L. & Ordal, G. W. (2004). Diversity in chemotaxis mechanisms among the Bacteria and Archaea. *Microbiol Mol Biol Rev* **68**, 301-319.

Thomas, C. M. (2000). Paradigms of plasmid organization. *Mol Microbiol* **37**, 485-491.

Thomas, C. M. & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 711-721.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876-4882.

Thorsted, P. B., Macartney, D. P., Akhtar, P., Haines, A. S., Ali, N., Davidson, P., Stafford, T., Pocklington, M. J., Pansegrau, W., Wilkins, B. M., Lanka, E. & Thomas, C. M. (1998). Complete sequence of the IncP beta plasmid R751: implications for evolution and organisation of the IncP backbone. *J Mol Biol* **282**, 969-990.

Timms-Wilson, T. (1998). *Molecular study and genetic approaches for improving the biological control activity of Pseudomonas fluorescens 54/96 in the suppression of Pythium ultimum in seedlings (Damping-off disease).* pp46. DPhil Thesis. Oxford University.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J.

D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M. & Venter, J. C. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539-547.

Top, E., De Smet, I., Verstraete, W., Dijkmans, R. & Mergeay, M. (1994). Exogenous isolation of mobilizing plasmids from polluted soils and sludges. *Appl Environ Microbiol* **60**, 831-839.

Top, E. M. & Springael, D. (2003). The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Curr Opin Biotechnol* **14**, 262-269.

Toussaint, A. & Merlin, C. (2002). Mobile elements as a combination of functional modules. *Plasmid* **47**, 26-35.

Turner, P. E., Cooper, V. S. & Lenski, R. E. (1998). Tradeoff between horizontal and vertical modes of transmission in bacterial plasmids. *Evolution* **52**, 315-329.

Turner, S. L., Bailey, M. J., Lilley, A. K. & Thomas, C. M. (2002). Ecological and molecular maintenance strategies of mobile genetic elements. *FEMS Microbiol Ecol* **42**, 177-185.

Turner, S. L., Lilley, A. K. & Bailey, M. J. (2002). Two *dna*B genes are associated with the origin of replication of pQBR55, an exogenously isolated plasmid from the rhizosphere of sugar beet. *FEMS Microbiol Ecol* **42**, 209-215.

van Elsas, J. D., Fry, J., Hirsch, P. & Molin, S. (2000). Ecology of Plasmid Transfer and Spread. In *The Horizontal Gene Pool: Bacterial plasmids and gene spread*, pp. 175-199. Edited by C. M. Thomas. Amsterdam: Harward academic publishers.

van Elsas, J. D., Turner, S. & Bailey, M. J. (2003). Horizontal gene transfer in the phytosphere. *New Phytol* **157**, 525-537.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H. & Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74.

Vetriani, C., Chew, Y. S., Miller, S. M., Yagi, J., Coombs, J., Lutz, R. A. & Barkay, T. (2005). Mercury adaptation among bacteria from a deep-sea hydrothermal vent. *Appl Environ Microbiol* **71**, 220-226.

Viegas, C. A., Lilley, A. K., Bruce, K. & Bailey, M. J. (1997). Description of a novel plasmid replicative origin from a genetically distinct family of conjugative plasmids associated with phytosphere microflora. *FEMS Microbiol Lett* **149**, 121-127.

Volker, U. & Hecker, M. (2005). From genomics via proteomics to cellular physiology of the Gram positive model organism *Bacillus subtilis*. *Cell Microbiol* **7**, 1077-1085.

Wang, M., Ahrne, S., Jeppsson, B. & Molin, G. (2005). Comparison of bacterial diversity along the human intestinal tract by direct cloning and sequencing of 16S rRNA genes. *FEMS Microbiol Ecol* **54**, 219-231.

Ward, N., Eisen, J., Fraser, C. & Stackebrandt, E. (2001). Sequenced strains must be saved from extinction. *Nature* **414**, 148-148.

Warren, G. J., Twigg, A. J. & Sherratt, D. J. (1978). ColE1 plasmid mobility and relaxation complex. *Nature* **274**, 259-261.

Weiss, A. A., Johnson, F. D. & Burns, D. L. (1993). Molecular characterization of an operon required for pertussis toxin secretion. *Proc Natl Acad Sci U S A* **90**, 2970-2974.

Wheatcroft, R. & Williams, P. A. (1981). Rapid methods for the study of both stable and unstable plasmids in *Pseudomonas*. *J Gen Microbiol* **124**, 433-437.

Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J. C., Yan, J. X., Gooley, A. A., Hughes, G., HumpherySmith, I., Williams, K. L. & Hochstrasser, D. F. (1996). From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (NY)* **14**, 61-65.

Willetts, N. & Crowther, C. (1981). Mobilization of the non-conjugative IncQ Plasmid RSf1010. *Genet Res* **37**, 311-316.

Williams, D. R. & Thomas, C. M. (1992). Active partitioning of bacterial plasmids. *J Gen Microbiol* **138**, 1-16.

Williams, H. G., Day, M. J., Fry, J. C. & Stewart, G. J. (1996). Natural transformation in river epilithon. *Appl Environ Microbiol* **62**, 2994-2998.

Wood, D. W., Setubal, J. C., Kaul, R., Monks, D. E., Kitajima, J. P., Okura, V. K., Zhou, Y., Chen, L., Wood, G. E., Almeida, N. F., Woo, L., Chen, Y. C., Paulsen, I. T., Eisen, J. A., Karp, P. D., Bovee, D., Chapman, P., Clendenning, J., Deatherage, G., Gillet, W., Grant, C., Kutyavin, T., Levy, R., Li, M. J., McClelland, E., Palmieri, A., Raymond, C., Rouse, G., Saenphimmachak, C., Wu, Z. N., Romero, P., Gordon, D., Zhang, S. P., Yoo, H. Y., Tao, Y. M., Biddle, P., Jung, M., Krespan, W., Perry, M., Gordon-Kamm, B., Liao, L., Kim, S., Hendrick, C., Zhao, Z. Y., Dolan, M., Chumley, F., Tingey, S. V., Tomb, J. F., Gordon, M. P., Olson, M. V. & Nester, E. W. (2001). The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* **294**, 2317-2323.

Xu, J. (2006). Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol Ecol* **15**, 1713-1731.

Yoshida, K., Kobayashi, K., Miwa, Y., Kang, C. M., Matsunaga, M., Yamaguchi, H., Tojo, S., Yamamoto, M., Nishi, R., Ogasawara, N., Nakayama, T. & Fujita, Y. (2001). Combined transcriptome and proteome analysis as a powerful approach to study genes under glucose repression in *Bacillus subtilis*. *Nucleic Acids Res* **29**, 683-692.

Young, J. P. W. & Wexler, M. (1988). Sym plasmid and chromosomal genotypes are correlated in field populations of *Rhizobium leguminosarum. J Gen Microbiol* **134**, 2731-2739.

Young, J. P., Crossman, L. C., Johnston, A. W., Thomson, N. R., Ghazoui, Z. F., Hull, K. H., Wexler, M., Curson, A. R., Todd, J. D., Poole, P. S., Mauchline, T. H., East, A. K., Quail, M. A., Churcher, C., Arrowsmith, C., Cherevach, I., Chillingworth, T., Clarke, K., Cronin, A., Davis, P., Fraser, A., Hance, Z., Hauser, H., Jagels, K., Moule, S., Mungall, K., Norbertczak, H., Rabbinowitsch, E., Sanders, M., Simmonds, M., Whitehead, S. & Parkhill, J. (2006). The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol* 7, R34.

Zechner, F., de la Cruz, F., Eisenbrandt, R., Grahn, A. M., Koraimann, G., Lanka, E., Muth, G., Pansegrau, W., Thomas, C. M., Wilkins, B. M. & Zatyka, M. (2000). Conjugative-DNA transfer processes. In *The Horizontal Gene Pool: Bacterial plasmids and gene spread*, pp. 87-174. Edited by C. M. Thomas. Amsterdam: Harward academic publishers.

Zhang, X. X., Lilley, A. K., Bailey, M. J. & Rainey, P. B. (2004)a. Functional and phylogenetic analysis of a plant-inducible oligoribonuclease (*orn*) gene from an indigenous *Pseudomonas* plasmid. *Microbiology (Reading, Engl)* **150**, 2889-2898.

Zhang, X. X., Lilley, A. K., Bailey, M. J. & Rainey, P. B. (2004)b. The indigenous *Pseudomonas* plasmid pQBR103 encodes plant-inducible genes, including three putative helicases. *FEMS Microbiol Ecol* **51**, 9-17.

Zhang, K., Martiny, A. C., Reppas, N. B., Barry, K. W., Malek, J., Chisholm, S. W. & Church, G. M. (2006). Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24**, 680-686.

Zhao, Y., Ma, Z. & Sundin, G. W. (2005). Comparative genomic analysis of the pPT23A plasmid family of *Pseudomonas syringae*. *J Bacteriol* **187**, 2113-2126.

Zhou, D., Han, Y., Song, Y., Tong, Z., Wang, J., Guo, Z., Pei, D., Pang, X., Zhai, J., Li, M., Cui, B., Qi, Z., Jin, L., Dai, R., Du, Z., Bao, J., Zhang, X., Yu, J., Wang, J., Huang, P. & Yang, R. (2004). DNA microarray analysis of genome dynamics in *Yersinia pestis*: insights into bacterial genome microevolution and niche adaptation. *J Bacteriol* **186**, 5138-5146.